Elucidation of Mechanisms of Fetal Hemoglobin Regulation by CRISPR/Cas9 Mediated Genome Editing

Elucidation of Mechanisms of Fetal Hemoglobin Regulation by CRISPR/Cas9 Mediated Genome Editing

A dissertation presented

by

Matthew Charles Canver

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

February 2016

Dissertation Advisor: Dr. Stuart H. Orkin Matthew Charles Canver

Elucidation of Mechanisms of Fetal Hemoglobin Regulation

by CRISPR/Cas9 Mediated Genome Editing

Abstract

Despite nearly complete understanding of the genetics of the β-hemoglobinopathies for several decades, definitive treatment options have lagged behind. Fetal hemoglobin (HbF) reinduction represents a “silver bullet” for therapy of the β-globin disorders. Recent development of the clustered regularly interspaced short palindromic repeat (CRISPR)/Cas9 nuclease system has allowed for facile manipulation of the genome for the study of genes and genetic elements. Here we developed CRISPR/Cas9-based methodology to reliably engender targeted genomic deletions ranging from 1.3 kilobases to over 1 megabase, which suggested an inverse relationship between deletion size and deletion frequency. Targeted deletion methods and Cas9-mediated *in situ* saturating mutagenesis were applied to the enhancer of the HbF repressor *BCL11A*, which revealed discrete vulnerabilities. This finding is consistent with emerging evidence in the field that large enhancers are comprised of constituent parts with some harboring the majority of the activity. The identified “Achilles heel” of the enhancer represents a promising therapeutic target. We further enhanced the resolution of the *in situ* saturating mutagenesis technique by using multiple Cas9 nucleases and variant-aware library design to identify functional sequences within the HBS1L-MYB intergenic region, a locus associated with elevated HbF levels. These data demonstrate the robustness of CRISPR/Cas9 mediated *in situ* saturating mutagenesis and targeted deletion to interrogate functional sequence within regulatory DNA. Harnessing the power of genome editing may usher in a second generation form of gene therapy for the β-globin disorders.

Elucidation of Mechanisms of Fetal Hemoglobin Regulation

by CRISPR/Cas9 Mediated Genome Editing

Table of Contents

Chapter 1

Introduction

The β-hemoglobinopathies, namely sickle cell disease (SCD) and β-thalassemia, result from genetic mutations in the β-globin gene and are among the most common monogenic diseases in the world[1]. SCD results from a nonsynonymous A to T mutation in codon 6 of the β-globin gene leading to a Glu-Val replacement[2,3] , whereas β-thalassemias are caused by diverse point mutations or deletions[4–9]. Treatment options are largely supportive. Transfusion and iron chelation are mainstays in the thalassemias while pain management, hydration, and hydroxyurea are used in SCD[10–16].

The hemoglobin tetramer is comprised of two α-like globin chains encoded by any of the three genes in the α-globin cluster on chromosome 16 and two β-like globin chains encoded from any of five genes in the β-globin locus on chromosome 11. The expression of the three genes at the α-globin locus (ζ, α_1, α_2) and the five genes at the β-globin locus (ε, $^{G}\gamma$, $^{A}\gamma$, δ, β) are developmentally regulated. It has been appreciated for many years that levels of fetal hemoglobin (HbF, $\alpha_2\gamma_2$), subject to developmental silencing in the months after birth, is a modifier of disease severity in patients with β-globin disorders[16–23]. This protective effect of HbF has motivated the therapeutic strategy to re-induce its expression in adult life. Hydroxyurea, a cytotoxic agent that inhibits ribonucleotide reductase, induces HbF modestly through an unknown mechanism. However, it has dose-limiting myelosuppressive effects and some patients are non-responders to therapy[10–13]. While bone marrow transplant represents the sole established curative option for patients, its use is limited by donor availability and graft versus host disease (GVH). A clinical trial has demonstrated successful gene addition of an anti-sickling form of β-globin to a transfusion-dependent $\beta^{E}\beta^{0}$ thalassemia patient that gained transfusion independence as a result of gene transfer[24]. Several additional somatic gene therapy trials for β-thalassemias and SCD are on-going[25]. Despite a deep understanding of

molecular defects and gene control mechanisms, treatment options for the majority of patients remain limited[3].

BREAKTHROUGHS IN GENOME EDITING TECHNOLOGIES

The emergence of designer nucleases for eukaryotic genome editing has prompted an era of unprecedented control over the genome. The development of zinc finger nucleases (ZFNs)[26–35], TAL effector nucleases (TALENs)[36–40], and meganucleases[41–44] established genome editing as a valuable laboratory technique. The emergence of the clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 nuclease system[45–49], which utilizes a single guide RNA (sgRNA) to direct the Cas9 nuclease for site-specific cleavage, has engendered tremendous excitement about potential clinical applications. The breakneck speed at which new variations on the general theme are developed is truly remarkable. Other Cas9-like systems include the CRISPR/Cpf1 nuclease platform[50], dimeric RNA-guided FokI nucleases[51,52], and use of Cas9's derived from a variety of prokaryotic species[53,54]. It is unlikely that the discovery of novel CRISPR-based systems and Cas9-like nucleases capable of eukaryotic genome editing will end soon[55].

The relative benefits of the newly developed CRISPR-based systems, ZFNs, and TALENs are still subject to debate. While CRISPR-based systems are often cited as the most efficient[56], ZFNs are the only editing technology that has been brought thus far to a clinical trial. The *CCR5* gene has been targeted by ZFNs in autologous $CD4^+$ T cells from patients with HIV. The gene-modified cells were subsequently reinfused, which led to a decrease in the blood level of HIV in most patients[57]. Notably, this study demonstrated that reinfusion of autologous genome edited primary human cells could be achieved, well-tolerated and possibly lead to clinical benefit.

Genome editing-based therapies rely on gene correction or disruption. Double strand break (DSB) induction by an engineered nuclease is repaired by the endogenous repair

pathways of homology-directed repair (HDR) or non-homologous end joining (NHEJ)[58]. Genetic correction strategies exploit the HDR pathway to insert custom sequences into the genome through co-delivery of an extrachromosomal repair template in conjunction with an engineered nuclease. The creation of a DSB improves HDR frequency[59]. As such, wild-type (or customized) sequences can be provided as an extrachromosomal donor for repair following site-specific cleavage by the nuclease. In contrast, genetic disruption strategies rely on the NHEJ pathway following nuclease-induced DSB to produce local insertions/deletions (indels)[47,48,58]. Introduction of two engineered nucleases can result in targeted deletion or inversion, duplication, local indels at nuclease cleavage sites, or translocations/chromosomal rearrangements[60–73].

ELUCIDATION OF MECHANISMS OF FETAL HEMOGLOBIN REGULATION BY CRISPR/CAS9 MEDIATED GENOME EDITING

In this dissertation work, I aimed to develop the ability to use CRISPR/Cas9 to engender targeted genomic deletions for the study of genes and genetic elements (Chapter 2). I also aimed to interrogate a genetic element necessary for HbF suppression through CRISPR-mediated targeted deletion and the novel approach of Cas9-mediated saturating mutagenesis (Chapter 3). I further aimed to develop computational methods to enhance the saturating mutagenesis technique through the usage of multiple nucleases and variant-aware library design (Chapter 4). Taken together, this dissertation work sought to use CRISPR/Cas9 genome editing to develop methodologies to study genes and genetics elements in erythroid cells to further understanding of HbF regulation and to identify novel therapeutic targets for HbF induction therapy for patients with β-globin disorders.

Chapter 2

Characterization of genomic deletion efficiency mediated by CRISPR/Cas9 in mammalian cells

ABSTRACT

The clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR associated (Cas) 9 nuclease system has provided a powerful tool for genome engineering. Double-strand breaks (DSBs) may trigger non-homologous end joining (NHEJ) repair, leading to frameshift mutations, or homology-directed repair (HDR) using an extrachromosomal template. Alternatively, genomic deletions may be produced by a pair of DSBs. The efficiency of CRISPR/Cas9-mediated genomic deletions has not been systematically explored. Here we present a methodology for the production of deletions in mammalian cells, ranging from 1.3 kilobases to greater than one megabase. We observed a high frequency of intended genomic deletions. Non-deleted alleles are nonetheless often edited with inversions or small indels produced at CRISPR recognition sites. Deleted alleles also typically include small indels at predicted deletion junctions. We retrieved cells with biallelic deletion at a frequency exceeding that of probabilistic expectation. We demonstrate an inverse relationship between deletion frequency and deletion size. This work suggests that CRISPR/Cas9 is a robust system to produce a spectrum of genomic deletions to allow investigation of genes and genetic elements.

INTRODUCTION

Recent studies have revealed a prokaryotic adaptive immune system which may be repurposed to allow for genomic engineering of eukaryotic genomes[45,46,48,74]. The *Streptococcus pyogenes* type II CRISPR/Cas9 adaptive immune system relies on three genes: two noncoding CRISPR RNAs (crRNAs) including a trans-activating crRNA (tracrRNA) and a precursor crRNA (pre-crRNA), as well as the CRISPR-associated Cas9 nuclease. The pre-crRNA is transcribed from an array that contains repetitive elements with interspersed unique sequences (spacers) derived

from exogenous DNA. Once processed and after interaction with the tracrRNA, the mature crRNA guides Cas9 to direct cleavage of foreign DNA[45,48,74,75]. This system has been repurposed for mammalian genome engineering using SpCas9 along with a fusion of the tracrRNA and mature crRNA to create a chimeric single guide RNA (sgRNA)[48,69,74]. Site-specific cleavage is directed by complementarity of the sgRNA to a 20-base pair genomic sequence (protospacer) immediately 5' of a protospacer-adjacent motif (PAM), which is NGG for Cas9. This recruits Cas9 to introduce site-specific DSBs repaired by either HDR or by insertion/deletion (indel)-forming NHEJ[48,74]. Heterologous expression of the CRISPR system components has been shown to be a facile method of genome engineering as compared to previous systems such as zinc finger nucleases (ZFNs) or TAL effector nucleases (TALENs) in part because of the ease with which individual sgRNAs may be designed and produced[56].

The CRISPR/Cas9 system has already demonstrated wide applicability for efficient genome editing in a variety of model systems[76–81], which has spawned an era of unprecedented control over the genome. This includes applications such as genome editing in clonal cell lines in a matter of weeks[69], CRISPR-interference (CRISPRi)-mediated gene regulation with a catalytically inactive Cas9[82], pooled sgRNA library screening for functional genomics[83–85], and potential CRISPR-based therapy highlighted by its recent use for gene correction in both murine and human stem cells[86,87].

A strategy of using two DSBs to create a deletion of the intervening segment by NHEJ has previously been successfully applied using ZFN, TALEN, and CRISPR systems[64–69,88]. However, the efficiency, reliability, and genomic outcomes of using pairs of CRISPRs to introduce genomic deletions remain incompletely characterized. Here we sought to test the capability and efficiency of creating deletions in mammalian cell lines. Our results indicate that the CRISPR/Cas9 system is a powerful tool for the robust and dependable generation of genomic deletions.

METHODS

CRISPR design and creation

sgRNA-specifying oligo sequences were chosen to minimize likelihood of off-target cleavage based on publicly available online tools[69]. Each sgRNA specified sequences exonic, intronic, or intergenic (within 3.5 kb of a gene body) with respect to a RefSeq gene (Figure 2.1a, b). "CACC" was added to the 5' end of the sgRNA-specifying oligo sequence and "AAAC" was added to the 5' end of the reverse complement of the sgRNA-specifying oligo for cloning using the BbsI restriction enzyme. G was added immediately following CACC if the first nucleotide was A, T, or C (in these cases C was added at the 3' end of the reverse complement oligo). The two oligos were phosphorylated and annealed using the following conditions: guide sequence oligo (10 μM), guide sequence reverse complement oligo (10 μM), T4 ligation buffer (1X) (New England Biolabs [NEB], Ipswich, MA), and T4 polynucleotide kinase (5 U) (NEB, Ipswich, MA) with the following temperature conditions: 37°C for 30 minutes; 95°C for 5 minutes and then ramp down to 25°C at 5°C/minute. The annealed oligos were cloned into pSpCas9(BB) (pX330; Addgene plasmid ID: 42230) using a Golden Gate Assembly strategy with the following conditions: 100 ng of circular pX330 vector, annealed oligos (0.2 μM), 2.1 buffer (1X) (NEB, Ipswich, MA), BbsI restriction enzyme (20 U) (NEB, Ipswich, MA), ATP (0.2 mM) (NEB, Ipswich, MA), BSA (1X) (NEB, Ipswich, MA), and T4 DNA ligase (750 U) (NEB, Ipswich, MA) with the cycling conditions of 20 cycles of 37°C for 5 minutes, 20°C for 5 minutes; 80°C for 20 minutes. The sgRNAs were not pre-screened for editing efficiency prior to genomic deletion experiments presented herein.

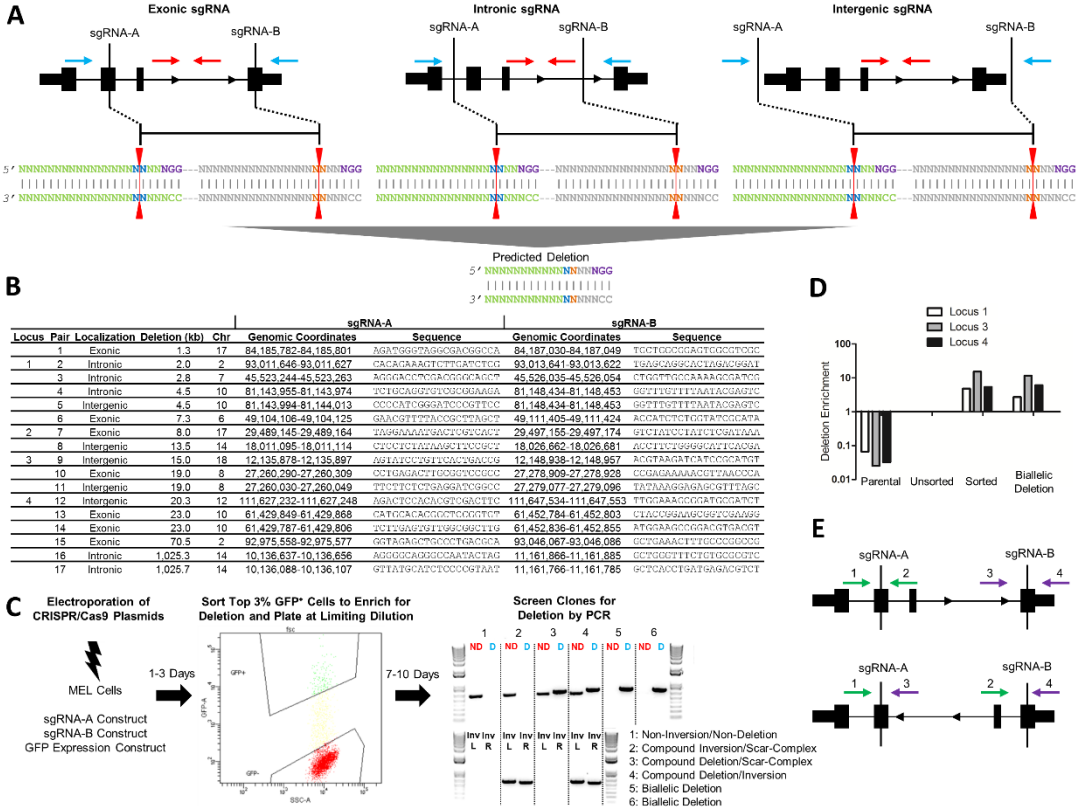
Cell culture, transfection, and screening clones

Murine erythroleukemia (MEL) cells were cultured in Dulbecco's modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 2% penicillin-streptomycin (PS), and 1% L-glutamine at 37°C with 5% CO_2 (Life Technologies, Green Island, NY). Despite a

complex karyotype, MEL cells exhibit karyotypic stability and are disomic for most chromosomes by karyotype reconstruction analysis[89,90]. Analysis of the MEL cells used in these experiments revealed a karyotype consistent with these previous reports. Two copies were present for each chromosome studied (data not shown). 2 x 10^6 cells were electroporated with 0.5 µg pmaxGFP plasmid (Lonza, Allendale, NJ) and 5 µg of each pX330-sgRNA plasmid using the ECM 830 Square Wave Electroporation System (Harvard Apparatus, Holliston, MA) for a total of 10.5 µg of plasmid[69]. Preliminary experiments showed this concentration of cells and plasmids to be effective for production of genomic deletions. Extensive analysis regarding optimal concentrations of cells and plasmids were not performed. Cells were resuspended in 100 µL BTX solution, and electroporated at 250 V, 5 ms, in 2 mm cuvettes (Harvard Apparatus, Holliston, MA). Cells were placed in 1 mL culture media immediately following electroporation.

To enrich for deletion, the top 3% GFP^+ cells were sorted via FACSAria cell sorter (BD Biosciences, San Jose, CA) 1-3 days post-electroporation (Figure 2.1c, d). Preliminary experiments showed these timepoints to be effective for production of genomic deletions. Extensive analysis regarding optimal timepoints for cell sorting were not performed. Cells were plated at 30 cells per 96-well plate to isolate single cell derived clones. After 7-10 days for expansion, clones were screened for CRISPR-mediated deletion (Figure 2.1c). Genomic DNA (gDNA) was extracted by resuspending cells in 50 µL QuickExtract DNA extraction solution per well and incubating at the following conditions: 65°C for 6 minutes; 98°C for 2 minutes (Epicentre, Madison, WI). Polymerase chain reaction (PCR) was performed using two sets of primers (Figure 2.1a, c): one set to amplify a sequence within the segment to be deleted (“non-deletion band”) and one set that only amplified in the presence of a deletion (“deletion band”) using the Qiagen HotStarTaq 2x master mix and the following cycling conditions: 95°C for 15 minutes; 35 cycles of 95°C for 15 seconds, 60°C for 1 minute, 72°C for 1 minute; 72°C for 10 minutes.

Monoallelic deletion clones were defined as having PCR amplification of both the non-deletion band and deletion band. Biallelic deletion clones were defined as having PCR



Locus	Pair	Localization	Deletion (kb)	Chr	sgRNA-A Genomic Coordinates	sgRNA-A Sequence	sgRNA-B Genomic Coordinates	sgRNA-B Sequence
	1	Exonic	1.3	17	84,185,782-84,185,801	AGATGGGTAGGGAAGGCCA	84,187,030-84,187,049	TCCTCCCGGAGTCGCCTCGC
1	2	Intronic	2.0	2	93,011,646-93,011,627	CACAGAAAGTCTTGATCTCG	93,013,641-93,013,622	TGAGCAGGCACTAGACGGAT
	3	Intronic	2.8	7	45,523,244-45,523,263	AGGGACCTCGAGGGCAGCT	45,526,035-45,526,054	CTGGTTGCCAAAAGCGATCC
	4	Intronic	4.5	10	81,143,955-81,143,974	TCTGCAGGTGTCGCGGAAGA	81,148,434-81,148,453	GGTTTGTTTCAATACGAGTC
	5	Intergenic	4.5	10	81,143,994-81,144,013	CCCCATCGGGATCCCGTTCC	81,148,434-81,148,453	GGTTTGTTTCAATACGAGTC
	6	Exonic	7.3	6	49,104,106-49,104,125	GAACGTTTTACCGCCTAGCT	49,111,405-49,111,424	ACCATCTCTGGTATCCCATA
2	7	Exonic	8.0	17	29,489,145-29,489,164	TAGAAAAATGACTCGTCACC	29,497,155-29,497,174	GTCTATCCTATCTCGATAAA
	8	Intergenic	13.5	14	18,011,095-18,011,114	CTCCTCTAAAAGCTTCCGCC	18,026,662-18,026,681	AGCTTCTGGGGCATTCACGA
3	9	Intergenic	15.0	18	12,135,878-12,135,897	AGTATCCTGTTCACCTGACCG	12,148,938-12,148,957	ACGTAAGATCATCCGCATGT
	10	Exonic	19.0	8	27,260,290-27,260,309	CCTGAGACTTGCGGTCCGCC	27,278,909-27,278,928	CCGAGAAAAACGTTAACCCA
	11	Intergenic	19.0	8	27,260,030-27,260,049	TTCTTCTCTGAGGATCGGCC	27,279,077-27,279,096	TATAAAGGAGAGCGTTTAGC
4	12	Intergenic	20.3	12	111,627,232-111,627,248	AGACTCCACACGTCGACTTC	111,647,534-111,647,553	TTGCAAACCCGATGCCATCT
	13	Exonic	23.0	10	61,429,849-61,429,868	CATCCACACCGCCTCCGTCT	61,452,784-61,452,803	CTACCGGAAAGCGGCCGAAGG
	14	Exonic	23.0	10	61,429,787-61,429,806	TCTTGAGTGTTGGTGGCTTG	61,452,836-61,452,855	ATGCAAGCCCGAACGTGAGT
	15	Exonic	70.5	2	92,975,558-92,975,577	GGTAGAGCTGCCCTGAGCCA	93,046,067-93,046,086	GCTGAAACTTTGCCCGGCCC
	16	Intronic	1,025.3	14	10,136,637-10,136,656	AGCGGCAGGGCCAATACTAG	11,161,866-11,161,885	GCTCGCTTTCTCTGCCGTTC
	17	Intronic	1,025.7	14	10,136,088-10,136,107	GTTATGCATCTCCCGTAAT	11,161,766-11,161,785	GCTCAGCTGATGAGACGTCT

Figure 2.1: Schema for CRISPR/Cas9-mediated genomic deletions. a, Exonic, intronic, and intergenic sgRNA deletion strategies. The sgRNA sites are shown in relation to an idealized gene. The PAM sequence (purple) is shown on the top strand for simplicity, but PAMs on both the top (Watson) and bottom (Crick) strands were used in different combinations. The red line indicates the predicted Cas9 cleavage between positions 17 and 18. The blue arrows indicate the position of PCR primers for deletion band amplification and the red arrows indicate the position of PCR primers for non-deletion band amplification. **b**, sgRNA localization (exonic/intronic/intergenic), deletion size, chromosome, genomic coordinates (mm10), and sequence for each sgRNA pair. Loci 1-4 used for further sequence analysis are indicated. **c**, CRISPR/Cas9-mediated genomic deletion strategy for MEL cells. 2 x 10^6 cells were electroporated with 5 µg of each sgRNA construct and 0.5 µg of a GFP expression construct. The top 3% GFP^+ cells were sorted 1-3 days post-electroporation and plated at limiting dilution. 7-10 days after plating, gDNA was extracted and clones screened for deletion by PCR. A representative screening agarose gel shows the detection of two non-deletion clones, two monoallelic deletion clones, and two biallelic deletion clones. The red "ND" refers to the non-deletion amplicon as schematized in A, and the blue "D" refers to the deletion amplicon as schematized in A. Upon inversion analysis, clones were further classified as non-deletion/non-inversion, compound inversion/scar-complex, compound deletion/scar-complex, and compound deletion/inversion. The distinction between scar and complex was established by presence or absence of PCR amplification flanking both sgRNA target recognition sites. "Inv" refers to inversion amplicons flanking left and right sgRNA recognition sites ("L" and "R", respectively). **d**, gDNA was extracted from cells prior to (unsorted) and after sorting the top 3% GFP+ cells (sorted). Deletion enrichment was calculated by RT-qPCR and data was normalized to the unsorted cells using the $2^{-\Delta Ct}$ method. A biallelic deletion clone for each locus was used as a positive control and non-edited parental gDNA as a negative control. **e**, top panel, Primers flanking the sgRNA recognition sites (shown in green and purple) were used to amplify 500-700 bp regions around each sgRNA site on non-deletion/non-inversion alleles (primers 1/2 and ¾, top panel). Inversion PCR utilized primer pairs (primers 1/3 and 2/4, respectively; bottom panel) in which both primers were in the same orientation, one inside and one outside the intended deletion.

amplification of the deletion band and absence of the non-deletion band. Clones with PCR amplification of the non-deletion band and absence of the deletion band were defined as non-deletion clones (Figure 2.1c).

Non-deletion/non-inversion alleles were analyzed using PCR primers flanking the sgRNA recognition sites to amplify 500-700 bp regions around each sgRNA site with one primer inside and one outside the intended deletion (Figure 2.1e, top panel). Monoallelic and biallelic inversion clones were defined by amplification of inversion bands at each inversion junction, using primer pairs in which both primers were in the same orientation, one inside and one outside the intended deletion (Figure 2.1e, bottom panel). Specifically, monoallelic inversion clones were defined as having PCR amplification using one or both inversion primer pairs. Biallelic inversion clones were defined as having PCR amplification of one or both inversion primer pairs in conjunction with neither deletion band PCR amplification nor amplification of sequences flanking the sgRNA recognition site.

Deletion, non-deletion, and inversion amplicons from non-deletion, monoallelic, and biallelic deletion clones were subjected to Sanger sequencing; the deletion amplicons from biallelic deletion clones were separately amplified by Phusion Hot Start PCR (New England Biolabs, Ipswich, MA) with the following conditions: Phusion High Fidelity DNA Polymerase (0.5 U), dNTPs (2 mM each), GC Buffer (1X), 3% dimethyl sulfoxide (DMSO), forward and reverse primers (0.1 μM each) with the cycling conditions of 95ºC for 15 minutes; 40 cycles of 95ºC for 15 seconds, 60ºC for 30 seconds, 72ºC for 30 seconds; 72ºC for 5 minutes. Amplicons were either gel purified and directly Sanger sequenced or cloned with Zero Blunt PCR Cloning kit (Life Technologies, Green Island, NY) where four clones each were subjected to Sanger sequencing. Deletion frequency was calculated as

$$100\%\left(\frac{(\#\,Monoallelic\ Deletions)\cdot(2\cdot\#\,Biallelic\ Deletions)}{2(\#\,Clones\ Screened)}\right).$$

Inversion frequency was calculated using the same equation.

RESULTS

CRISPR/Cas9 is a robust system for the production of genomic deletions

Seventeen sgRNA pairs at twelve genomic loci were assayed in MEL cells to determine their ability to engender genomic deletions and to determine the robustness of the approach. At each locus, a pair was comprised of an sgRNA 5' with respect to the top (Watson) strand, indicated as sgRNA-A, and another 3' with respect to the Watson strand, sgRNA-B (Figure 2.1a, b). The ability to create interstitial deletion of the segment AB was tested by conventional PCR. These sgRNAs were either exonic, intronic, or intergenic (within 3.5 kb of a gene body) with respect to RefSeq genes (Figure 2.1a, b). None of the genes was known to be essential for cell viability. The sgRNA pairs were designed to create a spectrum of deletions, ranging in size from 1.3 to 1,026 kb (Figure 2.1b). 1,974 clones across 17 sgRNA pairs were screened for deletions. Clones with deleted alleles were observed for all tested sgRNA pairs (Table 2.1).

Table 2.1: Observed biallelic deletion frequency exceeds probabilistic expectation. Expected (Exp) number of clones in each category (non-deletion clones, monoallelic deletion clones, biallelic deletion clones, and all clones) were calculated based on the observed (Obs) deletion frequency using the quadratic equation (analogous to Hardy-Weinberg equilibrium). In this case, monoallelic deletion clones consisted of compound deletion/scar, compound deletion/complex, and compound deletion/inversion clones. sgRNA pairs were analyzed individually and collectively for agreement with expected number of non-deletion, monoallelic, and biallelic deletion clones obtained. Chi square statistics (X^2) and p values were generated by Pearson's chi square tests. Inversion frequency was evaluated for the four pairs studied in detail as well as for pairs 16 and 17.

		Clones														Alleles					
		Non-Deletion				Monoallelic Deletion				Biallelic Deletion				All		Deletion			Inversion		
Deletion (kb)	Clones (n)	Obs	Exp	X^2	*p*	Obs	Exp	X^2	*p*	Obs	Exp	X^2	*p*	X^2	*p*	Del	Non-Del	Frequency	Inv	Non-Inv	Frequency
1.3	24	10	9	0.04	0.84	10	11	0.14	0.71	4	3	0.12	0.73	0.30	0.59	18	30	37.5%			
2	117	68	65	0.17	0.68	38	45	0.98	0.32	11	8	1.42	0.23	2.57	0.11	60	174	25.6%	20	136	12.8%
2.8	39	15	15	0.01	0.92	19	18	0.03	0.85	5	5	0.03	0.87	0.07	0.79	29	49	37.2%			
4.5	61	49	48	0.03	0.86	10	12	0.46	0.50	2	1	1.78	0.18	2.28	0.13	14	108	11.5%			
4.5	82	72	72	$1.29x10^{-3}$	0.97	10	9	0.04	0.84	0	0	0.30	0.58	0.35	0.56	10	154	6.1%			
7.3	166	115	112	0.07	0.79	43	49	0.63	0.43	8	5	1.45	0.23	2.15	0.14	59	273	17.8%			
8	400	242	233	0.38	0.54	126	145	2.46	0.12	32	23	3.95	$4.69x10^{-2}$	6.79	0.01	190	610	23.8%	9	87	9.4%
13.5	80	62	61	0.01	0.92	16	18	0.13	0.72	2	1	0.45	0.50	0.59	0.44	20	140	12.5%			
15	158	100	93	0.58	0.45	42	57	3.80	$5.13x10^{-2}$	16	9	6.21	$1.27x10^{-2}$	10.59	$1.14x10^{-3}$	74	242	23.4%	147	17	10.4%
19	34	32	32	$2.70x10^{-5}$	1.00	2	2	$1.78x10^{-3}$	0.97	0	0	0.03	0.86	0.03	0.86	2	66	2.9%			
19	120	103	100	0.10	0.76	13	19	1.98	0.16	4	1	10.33	$1.31x10^{-3}$	12.41	$4.27x10^{-4}$	21	219	8.8%			
20.3	70	42	40	0.09	0.77	22	26	0.54	0.46	6	4	0.85	0.36	1.48	0.22	34	106	24.3%	26	114	18.6%
23	71	55	52	0.13	0.72	12	17	1.56	0.21	4	1	4.77	0.03	6.46	$1.10x10^{-2}$	20	122	14.1%			
23	27	23	22	0.03	0.87	3	5	0.52	0.47	1	0	2.55	0.11	3.10	0.08	5	49	9.3%			
70.5	182	181	181	$1.04x10^{-9}$	1.00	1	1	$7.57x10^{-6}$	1.00	0	0	$1.37x10^{-3}$	0.97	$1.38x10^{-3}$	0.97	1	363	0.3%			
1025.3	133	132	132	$2.68x10^{-8}$	1.00	1	1	$1.42x10^{-5}$	1.00	0	0	$1.88x10^{-3}$	0.97	$1.89x10^{-3}$	0.97	1	265	0.4%	2	264	0.8%
1025.7	210	207	207	$5.55x10^{-7}$	1.00	3	3	$1.54x10^{-4}$	0.99	0	0	$1.07x10^{-2}$	0.92	$1.09x10^{-2}$	0.92	3	417	0.7%	2	416	0.5%
Total	**1,974**	**1,508**	**1,453**	**2.09**	**0.15**	**371**	**481**	**25.27**	$\mathbf{4.98x10^{-7}}$	**95**	**40**	**76.29**	$\mathbf{2.45x10^{-18}}$	**103.65**	$\mathbf{2.41x10^{-24}}$	**561**	**3,387**	**14.2%**	**76**	**1,164**	**5.8%**

Indels are often formed at the predicted deletion junctions

The deleted allele was examined in monoallelic and biallelic deletion clones at four loci, referred to as loci 1-4, spanning a range of intended deletion sizes (2.0, 8.0, 15.0, and 20.3 kb) (Figure 2.1b). One primer from upstream of sgRNA-A and another primer from downstream of sgRNA-B

were used to amplify and sequence across the deletion junction (Figures 2.1a, c; 2.2a, 2.3a, 2.4a, 2.5a). Indel sizes were enumerated as the sum of the number of base pairs inserted (positive values) or deleted (negative values) relative to the predicted deletions, assuming cleavage between positions 17 and 18 as specified by each sgRNA (Figure 2.1a) [48,69,74]. Indels ranged from -176 bp to +538 bp in monoallelic deletion clones and from -286 bp to +449 bp in biallelic deletion clones. Both monoallelic and biallelic deletion clones showed the preponderance of indels clustering between -10 to 0 bp (Figure 2.6a). Notably, absence of indels occurred as the most frequent outcome at each sgRNA predicted cleavage site in both monoallelic deletion clones (77/174 sites on 87 alleles from 87 clones, 44.3%) and biallelic deletion clones (17/80 sites on 40 alleles from 31 clones, 21.3%) (Figure 2.6a). Furthermore, the precise predicted deletion (i.e absence of indels at both sites on a deleted allele) occurred in 31.0% (27/87) of monoallelic deletion clones and 15.0% (6/40) of alleles from biallelic deletion clones. Positive indels (i.e. relative insertions to the predicted deletion) had homology to sequences flanking the predicted deletion site except in one case where a 538-bp insertion with sequence homology to a portion of the pX330 plasmid was identified. However, this accounting is likely an underestimate of the full spectrum of indels since the PCR-based screening method would not identify large deletions and insertions. In several instances, we identified large insertions and deletions around cleavage sites not detectable by the screening PCR strategy (data not shown).

Sanger sequencing of biallelic deletion clones revealed 9/31 (29.0%) to be compound heterozygotes based on differing indels at the predicted deletion sites on each allele. The remaining 22/31 (71.0%) had only a single deletion junction identified by both amplicon sequencing as well as sequencing of multiple clones of the PCR product. This finding may suggest that in these cells both alleles were independently repaired in an identical manner, or that one allele served as a template for HDR of the other allele. Other possibilities include uniparental disomy, monosomy, or a large insertion/deletion at one allele. Each of the biallelic

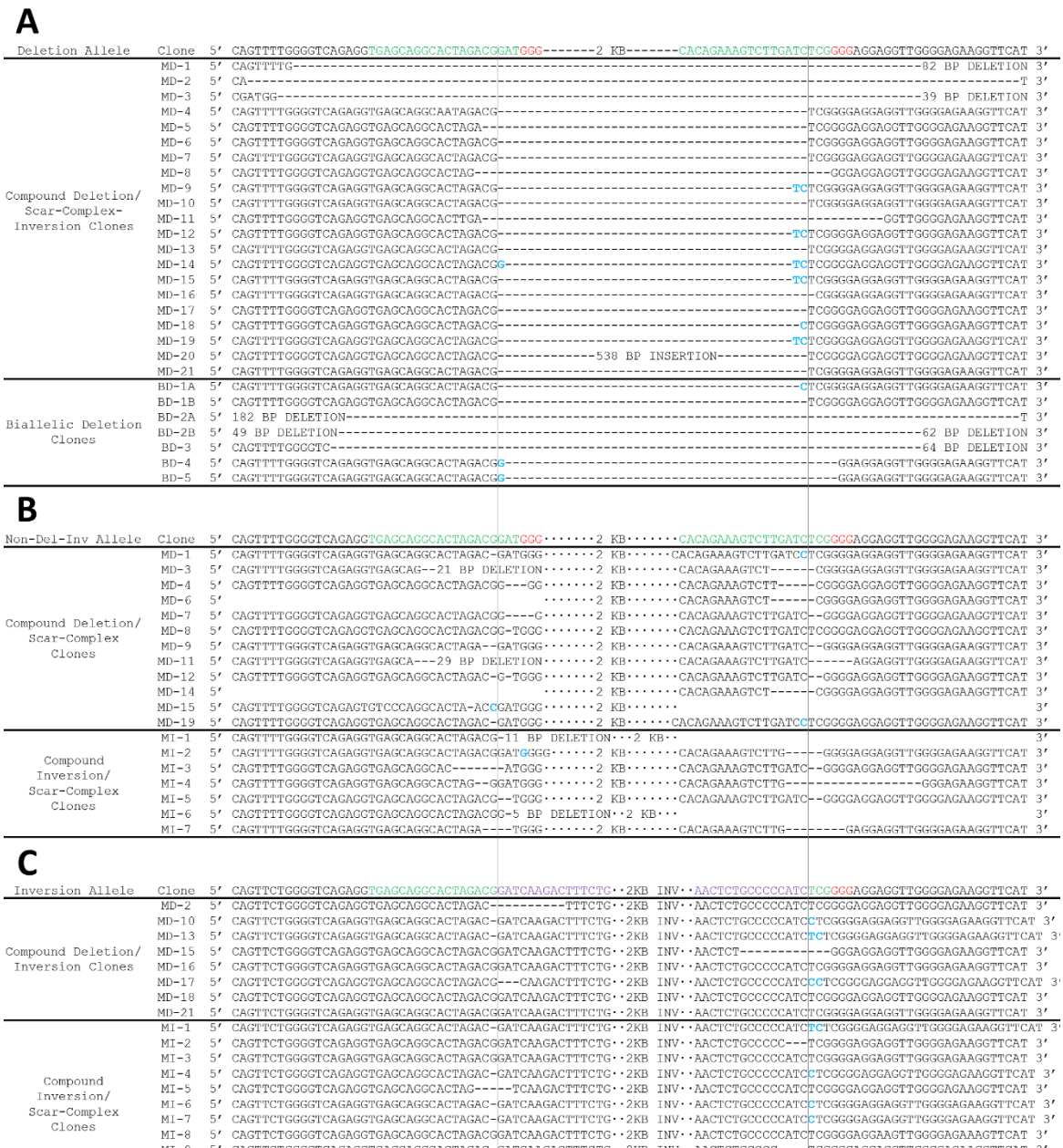


Figure 2.2: Indels at the non-deleted/non-inversion allele and indels at the predicted deletion/inversion junction at locus 1. a, Sequencing of the deletion allele in compound deletion/scar-complex-inversion clones and biallelic deletion clones at locus 1 (intended 2.0 kb deletion). Top row indicates sequence of unmodified allele. sgRNA sequences are shown in green and PAM sequences in red. Deletion events are shown by an equivalent number of dash marks and insertions are highlighted in blue. Vertical lines indicate predicted cleavage site, between positions 17 and 18 of the sgRNA. **b**, Sequencing of the non-deletion/non-inversion allele in compound deletion/scar-complex clones and compound inversion/scar-complex clones. Top row indicates sequence of the unmodified allele. sgRNA sequences are shown in green and PAM sequences in red. Deletion events are shown by an equivalent number of dash marks and insertions are highlighted in blue. Vertical lines indicate predicted cleavage site, between positions 17 and 18 of the sgRNA. **c**, Sequencing of the inversion allele in compound deletion/inversion clones and compound inversion/scar-complex clones. Top row indicates perfect inversion of intervening segment between predicted cleavage sites. sgRNA sequences are shown in green, PAM sequences in red, and inverted sequence in purple. Deletion events are shown by an equivalent number of dash marks and insertions are highlighted in blue. Vertical lines indicate predicted cleavage site, between positions 17 and 18 of the sgRNA. MD indicates monoallelic deletion, MI monoallelic inversion, and BD biallelic deletion.

Figure 2.3: Indels at the non-deleted/non-inversion allele and indels at the predicted deletion/inversion junction at locus 2. a, Sequencing of the deletion allele in compound deletion/scar-complex-inversion clones and biallelic deletion clones at locus 2 (intended 8.0 kb deletion). Top row indicates sequence of unmodified allele. sgRNA sequences are shown in green and PAM sequences in red. Deletion events are shown by an equivalent number of dash marks and insertions are highlighted in blue. Vertical lines indicate predicted cleavage site, between positions 17 and 18 of the sgRNA. **b**, Sequencing of the non-deletion/non-inversion allele in compound deletion/scar-complex clones and compound inversion/scar-complex clones. Top row indicates sequence of unmodified allele. sgRNA sequences are shown in green and PAM sequences in red. Deletion events are shown by an equivalent number of dash marks and insertions are highlighted in blue. Vertical lines indicate predicted cleavage site, between positions 17 and 18 of the sgRNA. **c**, Sequencing of the inversion allele in compound deletion/inversion clones and compound inversion/scar-complex clones. Top row indicates perfect inversion of intervening segment between predicted cleavage sites. sgRNA sequences are shown in green, PAM sequences in red, and inverted sequence in purple. Deletion events are shown by an equivalent number of dash marks and insertions are highlighted in blue. Vertical lines indicate predicted cleavage site, between positions 17 and 18 of the sgRNA. MD indicates monoallelic deletion, MI monoallelic inversion, and BD biallelic deletion. Clone MD10 exhibited amplification using PCR primers flanking sgRNA-site-B (see Figure 2.1e, top panel) and amplification using inversion PCR primers flanking sgRNA-site-A (see Figure 2.1e, bottom panel) suggesting a complex inversion. Clone MD2 exhibited amplification using PCR primers flanking sgRNA-site-A (see Figure 2.1e, top panel) and amplification using inversion PCR primers flanking both sgRNA-site-A and sgRNA-site-B (see Figure 2.1e, bottom panel) suggesting at least three copies at tested locus, which could be consistent with rare tetraploidies observed in MEL cells by karyotype (data not shown) or with a mixed clone. This was the only clone out of the 278 clones examined in detail across the four loci to exhibit apparent copy number greater than two.

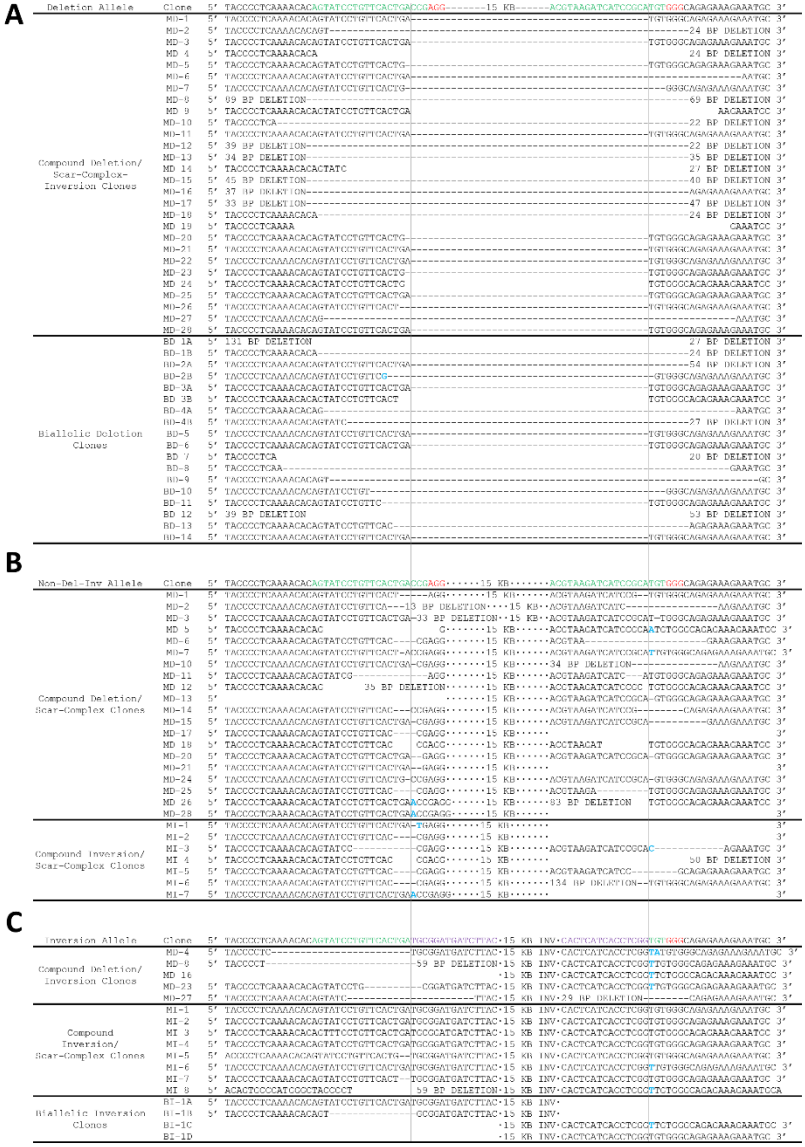


Figure 2.4: Indels at the non-deleted/non-inversion allele and indels at the predicted deletion/inversion junction at locus 3. a, Sequencing of the deletion allele in compound deletion/scar-complex-inversion clones and biallelic deletion clones at locus 3 (intended 15.0 kb deletion). Top row indicates sequence of unmodified allele. sgRNA sequences are shown in green and PAM sequences in red. Deletion events are shown by an equivalent number of dash marks and insertions are highlighted in blue. Vertical lines indicate predicted cleavage site, between positions 17 and 18 of the sgRNA. **b**, Sequencing of the non-deletion/non-inversion allele in compound deletion/scar-complex clones and compound inversion/scar-complex clones. Top row indicates sequence of unmodified allele. sgRNA sequences are shown in green and PAM sequences in red. Deletion events are shown by an equivalent number of dash marks and insertions are highlighted in blue. Vertical lines indicate predicted cleavage site, between positions 17 and 18 of the sgRNA. **c**, Sequencing of the inversion allele in compound deletion/inversion clones, compound inversion/scar-complex clones, and biallelic inversion clones. Top row indicates perfect inversion of intervening segment between predicted cleavage sites. sgRNA sequences are shown in green, PAM sequences in red, and inverted sequence in purple. Deletion events are shown by an equivalent number of dash marks and insertions are highlighted in blue. Vertical lines indicate predicted cleavage site, between positions 17 and 18 of the sgRNA. MD indicates monoallelic deletion, MI monoallelic inversion, BD biallelic deletion, and BI biallelic inversion. Biallelic inversion alleles are labeled A through D (rather than merely A and B) since phase of inversion junctions within a clone was not resolved. Clone MD16 did not exhibit amplification using PCR primers flanking either sgRNA-site-A or sgRNA-site-B (see Figure 2.1e, top panel), but did exhibit amplification using inversion PCR primers flanking sgRNA-site-B only (see Figure 2.1e, bottom panel) suggesting a complex inversion.

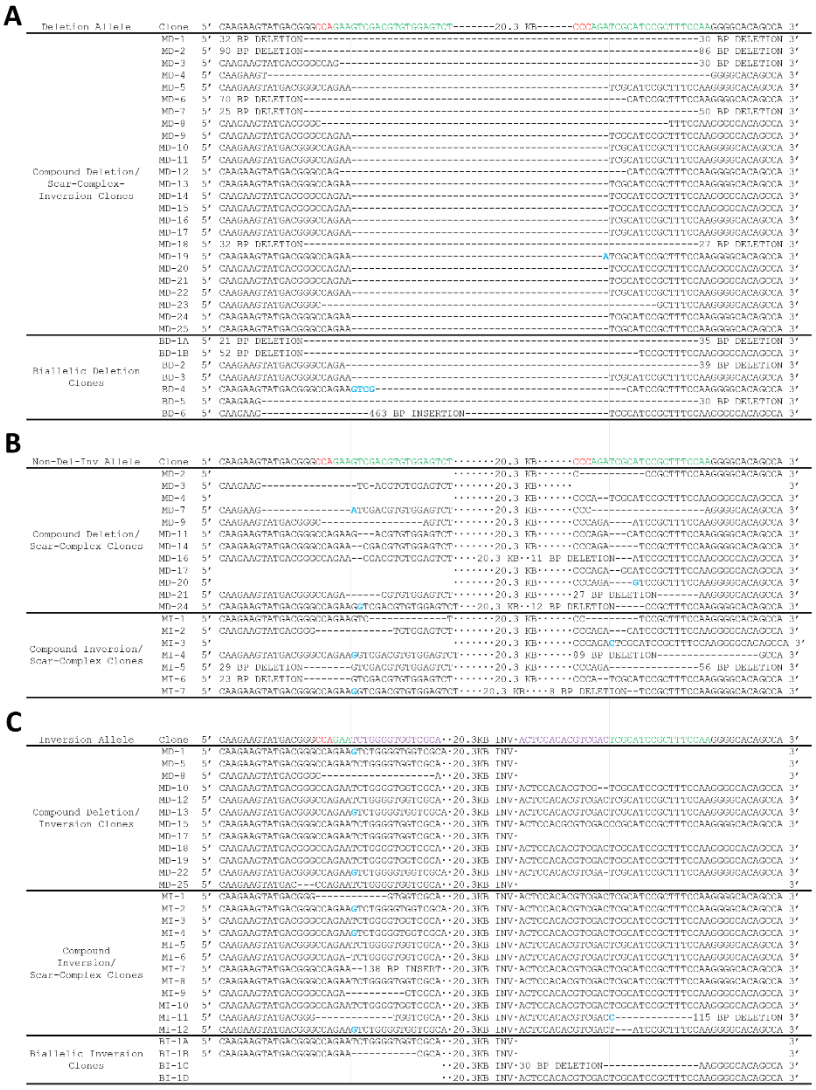


Figure 2.5: Indels at the non-deleted/non-inversion allele and indels at the predicted deletion/inversion junction at locus 4. a, Sequencing of the deletion allele in compound deletion/scar-complex-inversion clones and biallelic deletion clones at locus 4 (intended 20.3 kb deletion). Top row indicates sequence of unmodified allele. sgRNA sequences are shown in green and PAM sequences in red. Deletion events are shown by an equivalent number of dash marks and insertions are highlighted in blue. Vertical lines indicate predicted cleavage site, between positions 17 and 18 of the sgRNA. **b**, Sequencing of the non-deletion/non-inversion allele in compound deletion/scar-complex clones and compound inversion/scar-complex clones. Top row indicates sequence of unmodified allele. sgRNA sequences are shown in green and PAM sequences in red. Deletion events are shown by an equivalent number of dash marks and insertions are highlighted in blue. Vertical lines indicate predicted cleavage site, between positions 17 and 18 of the sgRNA. **c**, Sequencing of the inversion allele in compound deletion/inversion clones, compound inversion/scar-complex clones, and biallelic inversion. Top row indicates perfect inversion of intervening segment between predicted cleavage sites. sgRNA sequences are shown in green, PAM sequences in red, and inverted sequence in purple. Deletion events are shown by an equivalent number of dash marks and insertions are highlighted in blue. Vertical lines indicate predicted cleavage site, between positions 17 and 18 of the sgRNA. MD indicates monoallelic deletion, MI monoallelic inversion, BD biallelic deletion, and BI biallelic inversion. Biallelic inversion alleles are labeled A through D (rather than merely A and B) since phase of inversion junctions within a clone was not resolved. Clones MD1, MD5, MD8, and MD25 did not exhibit amplification using PCR primers flanking either sgRNA-site-A or sgRNA-site-B (see Figure 2.1e, top panel), but did exhibit amplification using inversion PCR primers flanking sgRNA-site-A only (see Figure 2.1e, bottom panel) suggesting complex inversions. Clone MD17 exhibited amplification using PCR primers flanking sgRNA-site-B (see Figure 2.1e, top panel) and amplification using inversion PCR primers flanking sgRNA-site-A (see Figure 2.1e, bottom panel) suggesting a complex inversion.

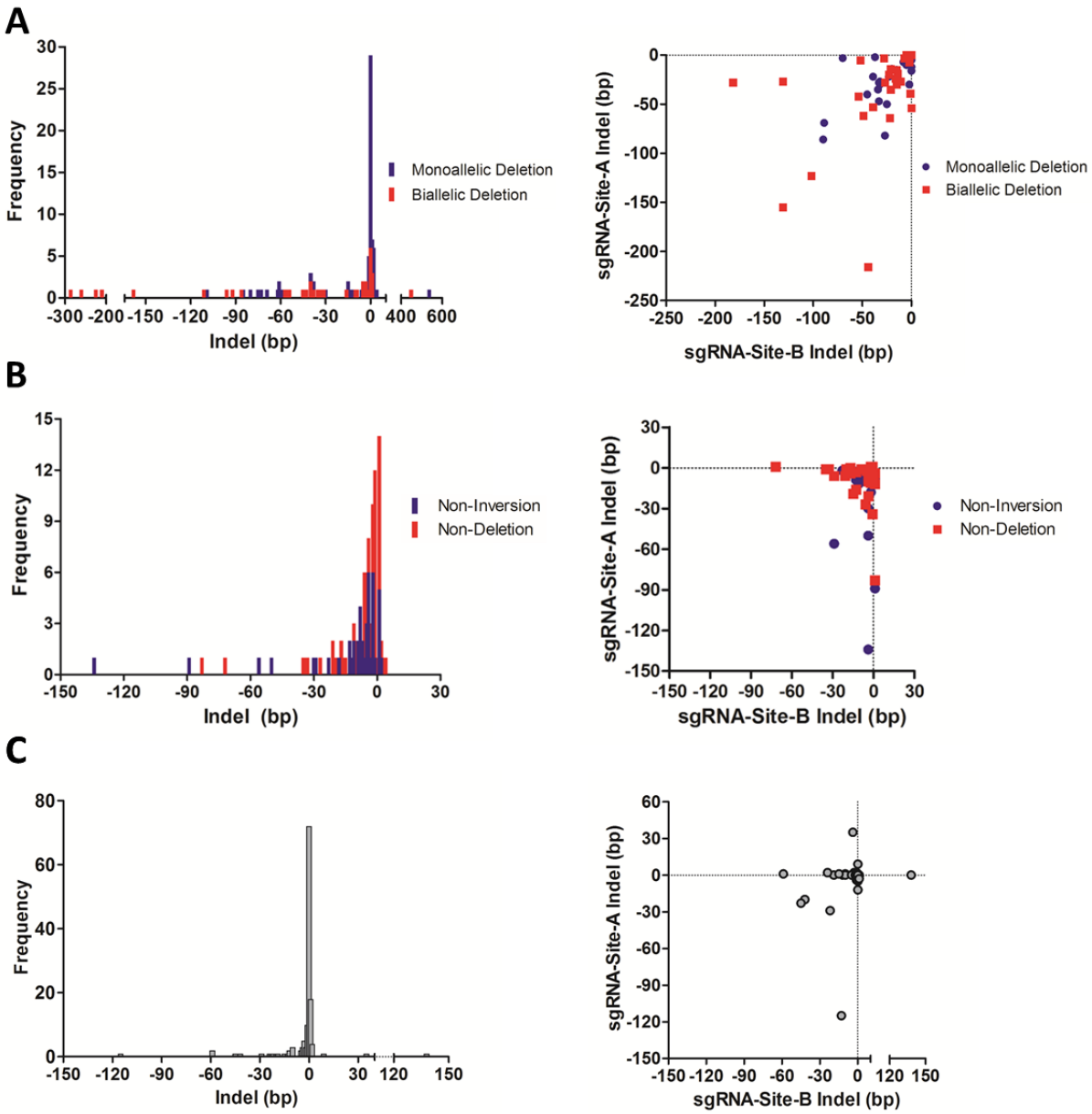


Figure 2.6: Indels at deleted, scarred, and inverted alleles. a, left panel, Frequency distribution of indel formation at the predicted deletion junctions from the deleted alleles of monoallelic and biallelic deletion clones across four loci examined in detail. Indels ranged from -176 bp to +538 bp in monoallelic deletion clones and from -286 bp to +449 bp in biallelic deletion clones with the majority found at -10 to 0 bp. **a**, right panel, Distribution of negative indels on the deletion allele produced by sgRNA-A and sgRNA-B from both monoallelic and biallelic deletion clones. Positive indels could not be reliably mapped to individual sites so were excluded from analysis. **b**, left panel, Frequency distribution of indel formation (scarring) on the non-deleted allele from compound deletion/scar-complex clones and on the non-inverted allele from compound inversion/scar-complex clones across eight sgRNA sites from four loci examined in detail. Indels ranged from -83 bp to +1 bp for the non-deletion alleles and from -134 bp to +2 bp for the non-inversion alleles. **b**, right panel, Distribution of indels on the non-deleted allele in compound deletion/scar clones and the non-inverted allele in compound inversion/scar clones in which sequences surrounding both sgRNA-A and sgRNA-B sites were amplified. **c**, left panel, Frequency distribution of indel formation at the predicted inversion junctions from the inversion alleles of compound inversion/scar-complex-deletion clones across all four loci examined in detail. Indels ranged from -115 bp to +138 bp with the majority found at -10 to 0 bp. **c**, right panel, Distribution of indels on the allele produced by sgRNA-A and sgRNA-B from compound inversion/scar-complex-deletion clones.

deletion clones was confirmed to result in loss of expression of the deleted gene by RT-qPCR (data not shown).

Editing ("scarring") occurs in the absence of deletion

Deletion is only one possible outcome following two DSBs at a locus. Another outcome is local NHEJ repair of each DSB without deletion of the intervening segment. We examined the non-deleted allele using Sanger sequencing of the monoallelic deletion clones from the same four sgRNA-pairs targeting different loci spanning a range of intended deletion sizes (2.0, 8.0, 15.0, and 20.3 kb) (Figures 2.2b, 2.3b, 2.4b, 2.5b). Primers flanking the sgRNA recognition sites were used to amplify 500-700 bp regions around each sgRNA site (Figure 2.1e, top panel). Three amplification patterns were observed for each sgRNA pair: amplification at both sgRNA-site-A and sgRNA-site-B (50/87, 57.5%), amplification at either sgRNA-site-A or sgRNA-site-B (11/87, 12.6%), or amplification at neither sgRNA-site-A nor sgRNA-site-B (26/87, 29.9%). For each locus, both sgRNA sites amplified from parental gDNA. The sites lacking amplification could represent large indels, inversions, or structural aberrancy and were excluded from this analysis. Clones with both sites amplifying were classified as "compound deletion/scar" clones while clones with one site or both sites not amplifying were classified as "compound deletion/complex" clones. Sanger sequencing of the amplifying sgRNA sites revealed indels at 99% (92/93) of eight sgRNA predicted cleavage sites across the four loci. Only one site exhibited either non-cleavage or cleavage with perfect repair (Figures 2.2b, 2.3b, 2.4b, 2.5b). No monoallelic deletion clones were obtained that exhibited a precise predicted deletion in combination with an absence of scarring of the non-deleted allele. The high frequency of indel formation across all four loci suggested that the frequency of indel formation was independent of the intended deletion size. Indels ranged from -83 bp to +1 bp with most of the indels clustering between -10 and -1 bp (Figure 2.6b). This high frequency of indel formation at the sgRNA sites on the non-deleted allele of monoallelic deletion clones indicates that the sgRNA recognition and/or PAM

sequence is often obliterated (Figures 2.2b, 2.3b, 2.4b, 2.5b, 2.6b). Therefore, re-targeting the non-deleted allele in monoallelic deletion clones would likely require independent sgRNAs.

From the non-deletion clones, 131/161 were chosen for analysis using the same primers flanking the sgRNA recognition sites to amplify 500-700 bp regions around each sgRNA site (Figure 2.1e, top panel). The same three amplification patterns were observed for each sgRNA pair: amplification at both sgRNA-site-A and sgRNA-site-B (110/131, 84.0%), amplification at either sgRNA-site-A or sgRNA-site-B (10/131, 7.6%), or amplification at neither sgRNA-site-A nor sgRNA-site-B (11/131, 8.4%). Of the 120 clones with at least one PCR amplification band, 108 clones were selected and subjected to Sanger sequencing (*n*=108). 24.1% (26/108) of these clones were subsequently determined to be monoallelic or biallelic inversion clones. Analysis of the remaining 82 sequenced non-deletion clones, classified as “non-deletion/non-inversion” clones, revealed indel formation of at least one allele in 75.6% (62/82) of instances with no apparent relationship between editing frequency and intended deletion size (data not shown). 24.4% (20/82) of the non-deletion/non-inversion clones exhibited wild-type sequencing, which could result from insufficient or absent sgRNA/Cas9 expression, perfect repair, or the presence of a large indel on one allele and an unmodified (or perfectly repaired) other allele. Based on the high frequency of indel formation, it is apparent that a large fraction of non-deletion clones were exposed to Cas9, sgRNA-A, and sgRNA-B. Therefore the induction of deletion is not simply limited by delivery of both CRISPR plasmids. However, the reduced rate of indel formation in non-deletion clones as compared to monoallelic deletion clones could indicate reduced delivery of one or both CRISPR plasmids to the cells, suggesting that deletion may be sensitive to sgRNA/Cas9 dose.

Inversion is a frequent outcome

The alleles from monoallelic deletion clones exhibiting amplification at neither sgRNA-site-A nor sgRNA-site-B (26/87, 29.9%) on the non-deletion allele (Figure 2.1e, top panel) were screened

for inversions by PCR (Figure 2.1e, bottom panel). Notably, 100% of these clones (26/26) demonstrated inversions. Each of the non-deletion clones previously analyzed (*n*=131) was evaluated for inversions. This analysis revealed 29.8% (39/131) of non-deletion clones to be monoallelic inversion clones and 1.5% (2/131) biallelic inversion clones. Notably, for the 8.4% (11/131) of non-deletion clones with PCR amplification at neither sgRNA-site-A nor sgRNA-site-B (11/131, 8.4%) (Figure 2.1e, top panel), 9/11 were determined to be monoallelic inversion clones and 2/11 were determined to be biallelic inversion clones by inversion PCR (Figure 2.1e, bottom panel). Sanger sequencing of the inversion amplicons revealed indels ranging from -115 bp to +138 bp for compound inversion/scar-complex-deletion clones at the predicted inversion junction (Figures 2.2c, 2.3c, 2.4c, 2.5c, 2.6c). This included a +35 bp insertion with homology to Cas9 and a +138 bp insertion with homology to a different chromosomal locus. Monoallelic inversion clones showed the preponderance of indels clustering between -10 to 0 bp (Figure 2.6c). Two biallelic inversion clones were identified and also exhibited small indel formation at the inversion junction (Figures 2.4c, 2.5c).

The non-inversion allele from the monoallelic inversion clones was analyzed for editing in the absence of inversion (Figure 2.1e, top panel). Sanger sequencing of the amplifying sgRNA sites revealed indels at 100% (45/45) of sgRNA cleavage sites analyzed, which ranged from -134 bp to +2 bp (Figures 2.2b, 2.3b, 2.4b, 2.5b, 2.6b). No clones were obtained that exhibited the predicted inversion in combination with an absence of scarring on the non-inversion allele. Of the monoallelic inversion clones, 59.0% (23/39) were “compound inversion/scar” clones and 41.0% (16/39) “compound inversion/complex” clones.

Pairs 16 and 17 each possess greater than 1 megabase between sgRNA-A and sgRNA-B (Figure 2.1b). Inversions were identified in 1.5% (2/133) and 1.0% (2/210) of clones, respectively (Table 2.1).,These data indicate that both large scale deletions and inversions of at least one megabase can be produced using two sgRNA.

Deletion occurs more frequently than inversion

The four loci analyzed in detail include a total of 278 clones (558 alleles). Deletion and inversion frequencies were calculated on a per allele basis, which revealed a deletion frequency of 26.8% (149/558) and an inversion frequency of 12.9% (72/558) (Figure 2.7a). Alleles characterized by non-editing, scarring, and complex indels remained the most common outcome (60.3%, 335/556). Clones were classified into 8 categories based on combination of deletion, inversion, scar/non-edited, and complex alleles. The assignment of scarring required PCR amplification of both the sgRNA-A and sgRNA-B target sites from the non-deletion or non-inversion allele. If one or both sites failed to amplify, alleles were classified as complex. Non-deletion/non-inversion clones were the most frequent outcome (42.8%, 119/278). Monoallelic deletion with scarring on the non-deletion allele was second most common (14.7%, 41/278) (Figure 2.7b). While biallelic inversion was least common (0.7%, 2/278), biallelic deletion was third most common (11.2%, 31/278). This distribution may reflect the complicated repair processes cells undergo after a pair of DSBs initiated by CRISPR/Cas9.

Observed frequency of monoallelic and biallelic deletion clones deviates from probabilistic expectation

A quadratic equation (analogous to Hardy-Weinberg equilibrium) was used to calculate the expected number of monoallelic and biallelic deletion clones for each deletion based on the observed deletion frequency:

$$f_{Non-Deletion}^2 + 2f_{Non-Deletion}f_{Deletion} + f_{Deletion}^2,$$

where f represents allele frequency. A Pearson's chi square test demonstrated that the observed frequency of non-deletion, monoallelic and biallelic deletion clones did not match expectations (p=2.41x10^{-24}). The deviation from probabilistic expectations occurred due to an increased frequency of biallelic deletion clones at the expense of the observed number of monoallelic deletion clones (Table 2.1). This suggests that the deletion of one allele may

increase the likelihood of deletion of the other allele. This finding could be consistent with the hypothesis that high expression of components of the CRISPR/Cas9 system favor production of genomic deletions. In any event, this observation is favorable for the creation of biallelic deletions for the study of genes or regulatory elements.

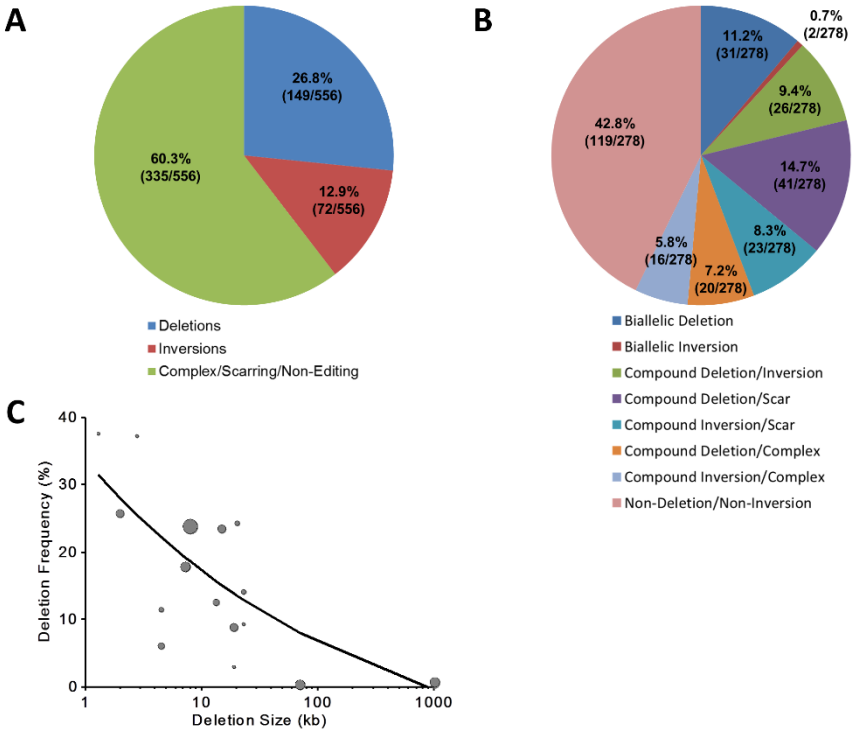


Figure 2.7: Deletion/inversion frequency and characterization of outcomes when using two sgRNA. a, Deletion and inversion frequency were calculated on a per allele basis for all 278 clones (556 alleles) across the four loci examined in detail. **b**, The 278 clones from across the four loci examined in detail were classified into eight categories based on presence of deletion, inversion, scar, and complex alleles. **c**, Deletion frequency inversely related to deletion size. Genomic deletion sizes ranged from 1.3 kb to 1,026 kb, which show a decrease in the frequency of deletion events as deletion size increases. The size of the circles corresponds to the number of clones screened for the corresponding sgRNA pair. The best fit relationship was determined by a weighted (by number of clones screened) non-linear regression of the form function(Deletion Size) = $k_1 + k_2(\text{Deletion Size})^{-k_3}$, where k_1, k_2, and k_3 represent constants (function(Deletion Size) = $-2.84+41.41(\text{Deletion Size})^{-0.36}$; $R^2 = 0.62$). The weighted non-linear regression was computed using the fitnlm function available in MATLAB R2013b software (MathWorks, Natick, MA).

Deletion frequency varies inversely to deletion size

Genomic deletion demonstrated an inverse relationship between genomic deletion size and frequency of deletion over the range of 1.3 to 1026 kb (Figure 2.7c). The best fit relationship was determined by a weighted (by number of clones screened) non-linear regression of the form function(Deletion Size) = $k_1 + k_2(\text{Deletion Size})^{-k_3}$, where k_1, k_2, and k_3 represent constants

(function(Deletion Size) = -2.84+41.41(Deletion Size)$^{-0.36}$; R^2 =0.62). The weighted non-linear regression was computed using the *fitnlm* function available in MATLAB R2013b software (MathWorks, Natick, MA). These data appear consistent with a competitive relationship between exposed cleaved ends at sgRNA sites-A and -B for local A-A indel repair and long-range A-B deletion repair (as well as inversion repair). Practically, these results suggest that more clones need to be screened as intended deletion size increases to reliably retrieve biallelic deletion clones.

DISCUSSION

Efficient genomic deletion may be useful for creation of specific and reproducible gene knockouts as compared to those produced by a single sgRNA to rely on indel-forming NHEJ to produce frameshift mutations. First, only 2/3 indels would result in a frameshift. Second, many frameshift mutations escape nonsense-mediated decay[91]. Third, alternative splicing may result in isoforms that may compensate or change gene function.

Given the high observed frequency of scarring of the non-deleted allele and the inverse relationship between deletion size and frequency, we suggest that two exonic sgRNAs designed to create a disruptive interstitial deletion of modest size (see Figure 2.1a, left panel) could be an efficient strategy to produce gene knockout clones. Rapid screening via PCR would identify clones with an appropriate deletion. Even in the event that adequate biallelic deletion clones could not be easily obtained, a monoallelic deletion would enrich for compound heterozygous loss-of-function alleles, which could be confirmed by measurement of RNA and protein levels.

Off-target cleavages from CRISPR/Cas9 have been reported, particularly at sites of sequence match in the PAM-proximal seed region[92–94]. The full extent of CRISPR/Cas9-mediated off-target events such as might be determined by deep whole genome sequencing remains incompletely characterized. The introduction of two sgRNAs as described in our deletion schema theoretically carries twice the frequency of off-target mutations as compared to

a single sgRNA. Off-target effects may be minimized using a double nickase strategy for DSBs[68,69], truncated sgRNAs[95], or dimeric Cas9-FokI fusions[51,96] but it remains to be determined how these approaches might affect deletion frequency. Complementation of the deletion phenotype by reintroduction of the deleted sequence would verify the association of the deletion to the phenotype but may be laborious to achieve. Alternatively, we suggest that multiple sgRNA pairs with non-overlapping sequences be utilized as a practical measure to control for off-target effects. Consistent phenotypes associated with deletions produced by unique reagents would seem unlikely due to off-target mutations.

These results identify an inverse relationship between deletion size and frequency. Of course a chromosome is not organized as a purely one-dimensional structure but rather as three-dimensional chromatin. The relationship between genome editing and chromatin context remains poorly understood and is an important topic for future investigation. The loci analyzed in this study mainly represent euchromatin. Although we observed some variation in deletion efficiency at different loci, based on a given deletion size a minimal number of clones needed to be screened to obtain biallelic deletion clones may be estimated. If biallelic deletion frequency was found below probabilistic expectation, it could suggest that full loss-of-function is cell lethal. Alternatively, lower than expected apparent biallelic deletion frequency in conjunction with the presence of apparent monoallelic deletion clones could indicate gene copy number greater than two, particularly in cancer cell lines.

These data suggest that both genomic deletions and inversions are a common outcome of a pair of DSBs initiated by CRISPR/Cas9. Despite relatively high deletion and inversion frequency, the paucity of biallelic inversion as compared to biallelic deletion clones may reflect a more complicated mechanism of inversion resolution as compared to deletion production. It is possible that some of the excess of biallelic deletion clones indicate circumstances in which the first deletion allele served as template for HDR of the second allele. For pairs of sgRNAs separated by greater than 1 Mb, we observed a similar frequency of deletions and inversions

(0.58% or 4/686 for both inversions and deletions; Figure 2.1b, Table 2.1). These results suggest that this technique may be suitable for the production of even relatively large intrachromosomal inversions and deletions.

The high scarring frequency observed related to indel production at the sgRNA recognition sites on non-deleted/non-inverted alleles provides insight into the efficiency of the CRISPR/Cas9 system. Even transient transfection of components of this system, can induce extensive on-target editing in the form of deletions, inversions, and scarring (i.e. small indels) in selected subpopulations. For example, it would be difficult to retarget monoallelic deletion clones with the same sgRNAs since in most cases the recognition sites have been altered by indels. Although small indels at deletion and inversion junctions are common, we also observed alleles with the precise predicted deletion or inversion without additional indels.

It has been previously observed that individual sgRNAs in the presence of Cas9 may show differences in editing efficiency[95]. We observed relatively substantial editing efficiencies for all tested sgRNA pairs (Table 2.1). This finding could indicate that enrichment via marker-based sorting can minimize differences in sgRNA efficiency, although the efficiency of these same sgRNA pairs to produce deletion without marker-based sorting was not characterized in detail.

The CRISPR-mediated deletion strategy appears particularly suited for the study of non-coding regulatory DNA where frameshift mutations do not pertain[67]. The importance of efficient strategies for assessment of regulatory DNA function is emphasized by recent reports that have suggested the majority of variants associated with human disease reside in non-coding regulatory DNA[97]. Combinatorial deletion of regulatory elements and/or genes may be a powerful method for studying pathways, the hierarchy of regulatory elements, and gene-regulatory element interactions. This study demonstrates that the CRISPR/Cas9 system is a robust tool to produce biallelic genomic deletions for prospective functional genetics.

ACKNOWLEDGMENTS

This text has been reproduced from the following article published in the *Journal of Biological Chemistry*: **Canver, MC**., Bauer, DE., Dass, A., Yien, YY., Chung, J., Masuda, T., Maeda, T., Paw, BH, Orkin, SH. (2014). Characterization of Genomic Deletion Efficiency Mediated by Clustered Regularly Interspaced Palindromic Repeats (CRISPR)/Cas9 Nuclease System in Mammalian Cells. *Journal of Biological Chemistry*. 289(31): 21312-24

Thanks to Matthew Porteus for suggesting genomic deletions by genome editing as a strategy to produce loss-of-function alleles, Jason Wright for suggesting the Golden Gate Assembly cloning strategy, DFCI FACS facility for cell sorting, Lisa Livanos (from KaryoLogic, Research Triangle Park, NC) for karyotype analysis, and members of the Orkin laboratory, particularly Jian Xu, Guoji Guo, Elenoe C. Smith, and Partha Das, for helpful discussions and sharing results. This work was supported by NIH R01HL032259 and P30DK049216 (Center of Excellence in Molecular Hematology) to S.H.O., NIDDK K08DK093705 to D.E.B, NIAID R01AI084905 to T.M., T32HL007574 and F32DK098866 to Y.Y.Y., R01DK070838 and P01HL032262 to B.H.P., and by CIHR fellowship to J.C.

M.C.C. conceived this study. M.C.C., D.E.B., A.D., Y.Y.Y., J.C., T.M, T.M., B.H.P, and S.H.O. performed or supervised clone screening analysis. M.C.C. acquired clone sequencing data. M.C.C., D.E.B., and S.H.O. analyzed the clone and sequencing data. M.C.C., D.E.B., and S.H.O. wrote the manuscript with input from all authors.

Chapter 3

BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis

ABSTRACT

Enhancers are critical determinants of cellular identity, yet their underlying structure remains poorly understood. Enhancers are commonly identified by correlative chromatin marks and gain-of-function potential, though only loss-of-function studies can demonstrate their requirement for gene expression in the native genomic context. Previously we identified an erythroid enhancer of *BCL11A*, subject to common genetic variation associated with fetal hemoglobin (HbF) level, whose mouse ortholog is necessary for erythroid BCL11A expression. Here we develop pooled CRISPR-Cas9 guide RNA libraries to perform *in situ* saturating mutagenesis of the human and mouse enhancers. This approach reveals critical minimal features and discrete vulnerabilities of these enhancers. Despite conserved function of the composite enhancers, their architecture diverges. The crucial human sequences appear primate-specific. Through editing of primary human progenitors and mouse transgenesis, we validate the BCL11A erythroid enhancer as a target for HbF reinduction. This detailed enhancer map will inform therapeutic genome editing.

INTRODUCTION

Enhancers are classically described as distal genetic elements that positively regulate gene expression in an orientation-independent manner in ectopic heterologous gain-of-function expression experiments[98]. These elements coordinate when, where, and how genes are expressed. Enhancer sequences bind transcription factors and are correlated with specific chromatin features including reduced DNA methylation, characteristic histone modifications, heightened chromatin accessibility, long-range promoter interactions, and bidirectional transcription. Recent chromatin mapping has demonstrated the abundance of distal regulatory elements bearing an enhancer signature[99–101].

The biological importance of enhancers is underscored by gene expression studies showing the predictive power of enhancer profile on lineage-specific programs[102–104]. Highly marked and clustered enhancers (e.g. so-called strong, stretch, or super-enhancers) are particularly suggestive of cellular identity and may help to infer lineage-specific regulatory factors[105–107]. Genome-wide association studies reveal enrichment of trait-associated variants in sequences bearing lineage-restricted enhancer signatures[97,101,105,108]. Enhancers display signs of evolutionary constraint as well as heightened turnover with evidence of positive selection[109–112].

Despite their importance, enhancers are typically defined by criteria unrelated to *in situ* functional requirement. Advances in putative enhancer mapping, as well as large-scale oligonucleotide synthesis, facilitate enhancer reporter assays on a massively parallel scale, allowing a systematic evaluation of the functional significance of enhancer sequences[113,114]. Nonetheless, ectopic heterologous enhancer assays cannot address the necessity of an element in its native chromatin environment. The growing appreciation of the nonrandom distribution of distal elements both with respect to the linear genome and within the three-dimensional nuclear environment emphasizes the importance of studying enhancers by perturbing their endogenous condition[107,115].

Insightful observations have been made by mutagenizing enhancers using traditional molecular genetic approaches[116,117]. However the low throughput of these classical methods constrains their widespread application. Furthermore the elevated turnover of many enhancer sequences between species may limit the ability to derive conclusions from nonhuman organisms regarding human gene regulation. Advances in genome editing technology make practical the facile modification of the human genome[47,48]. High-throughput Cas9-mediated functional genomics studies have revealed novel genes required for various biologic processes[83,84,118,119]. Genome editing is likewise suitable for the study of non-coding genetic

elements such as enhancers, although these experiments have previously been conducted at low-throughput[67,120,121].

METHODS

Generation of genomic deletions in HUDEP-2 cells

HUDEP clone 2 (HUDEP-2), was utilized as previously described[122]. HUDEP-2 cells were expanded in StemSpan SFEM (Stem Cell Technologies) supplemented with 10^{-6} M dexamethasone (Sigma), 100 ng/mL human stem cell factor (SCF) (R&D), 3 IU/mL erythropoietin (Amgen), 1% L-glutamine (Life Technologies), and 2% penicillin/streptomycin. 1 µg/mL doxycycline (Sigma) was included in the culture to induce expression of the human papilloma virus type 16 E6/E7 genes[122]. HUDEP-2 cells were differentiated in Iscove's Modified Dulbecco's Medium (IMDM) (Life Technologies) supplemented with 330 µg/mL holo-transferrin (Sigma), 10 µg/mL recombinant human insulin (Sigma), 2 IU/mL heparin (Sigma), 5% human solvent detergent pooled plasma AB (Rhode Island Blood Center), 3 IU/mL erythropoietin, 100 ng/mL human SCF, 1 µg/mL doxycycline, 1% L-glutamine, and 2% penicillin/streptomycin.

Tandem sgRNA lentiviruses were transduced into HUDEP-2 with stable Cas9 expression (Supplemental Table 3.1). Bulk cultures were incubated for 7-10 days with 10 µg/mL blasticidin and 1 µg/mL puromycin selection to allow for editing. Then bulk cultures were plated clonally at limiting dilution. 96 well plates with greater than 30 clones per plate were excluded to avoid mixed clones. After approximately 14 days of clonal expansion, genomic DNA was extracted using 50 µL QuickExtract DNA Extraction Solution per well (Epicentre). Clones were screened for deletion by conventional PCR with one PCR reaction internal to segment to be deleted ('non-deletion band') and one gap-PCR reaction across the deletion junction ('deletion band') that would only amplify in the presence of deletion[60]. Biallelic deletion clones were identified as the absence of the non-deletion PCR band and the presence of the deletion PCR band. Inversion clones were identified as previously described by PCR[60] (Supplemental Table

3.3). Briefly inversion clones had one inverted allele and one deleted allele without the presence of nondeletion alleles. In our experience biallelic inversion clones are very rare events[60]. PCR was performed using the Qiagen HotStarTaq 2x master mix and the following cycling conditions: 95°C for 15 minutes; 35 cycles of 95°C for 15 seconds, 60°C for 1 minute, 72°C for 1 minute; 72°C for 10 minutes. Alternatively, PCR was also performed using 2x Accuprime Supermix II (Life Technologies) with the following cycling conditions: 94°C for 2 minutes; 35 cycles of 94°C for 20 seconds, 60°C for 20 seconds, 68°C for 1 min/kb of PCR product; 68°C for 5 minutes. RNA was extracted from each positive clone using a kit (Qiagen) and quantitative real-time RT-qPCR was performed using iQ SYBR Green Supermix (Bio-Rad). Primers used are found in Supplemental Table 3.5. Gene expression was normalized to that of GAPDH. We isolated four control, one *BCL11A* null, three composite enhancer deleted, one h+55 deleted, one h+58 deleted, five h+62 deleted, three h+55 inverted, and two h+58 inverted clones. The *BCL11A* null clone had a 216 bp interstitial deletion of exon 2, preventing binding of the RT-qPCR primers. All gene expression data reported from these clones represents the mean of at least three technical replicates.

Design and synthesis of human and mouse lentiviral sgRNA libraries

Every 20-mer sequence upstream of an NGG or NAG PAM sequence on the plus or minus strand was identified for both the human and mouse orthologous +55, +58, and +62 DNase I hypersensitive site (DHS) as well as *BCL11A/Bcl11a* exon 2 (Supplemental Data). Relative to the human hg19 reference genome, a reference was used with the following substitutions to approximate a common low-HbF associated haplotype: rs1427407-G, rs1896293-T, rs6706648-T, rs6738440-G, rs7606173-C. The mouse orthologous sequences to each of the human DHSs were defined by using the liftOver tool of UCSC Genome Browser as previously described[67]. Each of the sgRNA oligos were synthesized as previously described[84,123,124] and cloned using a Gibson Assembly master mix (New England Biolabs) into lentiGuide-Puro (Addgene plasmid ID

52963) which had been BsmBI digested, PCR purified, and dephosphorylated. Gibson Assembly products were transformed to electrocompetent cells (E. cloni, Lucigen). Sufficient colonies were isolated to ensure ~90X library coverage for both human and mouse libraries. Plasmid libraries were deep sequenced to 533X and 813X coverage for human and mouse libraries respectively to confirm representation.

To produce lentivirus, HEK293T cells were cultured with Dulbecco's Modified Eagle's Medium (DMEM) (Life Technologies) supplemented with 10% fetal bovine serum (FBS) (Omega Scientific) and 2% penicillin-streptomycin (Life Technologies) in 15 cm tissue culture treated petri dishes. HEK293T were transfected at 80% confluence in 12 mL of media with 13.3 µg psPAX2, 6.7 µg VSV-G, and 20 µg of the lentiviral construct plasmid of interest using 180 µg of linear polyethylenimine (Polysciences). Medium was changed 16-24 hours after transfection. Lentiviral supernatant was collected at 48 and 72 hours post-transfection and subsequently concentrated by ultracentrifugation (24,000 rpm for 2 hours at 4ºC with Beckman Coulter SW 32 Ti rotor).

Tiled pooled CRISPR-Cas9 screen for *in situ* functional mapping the human *BCL11A* erythroid enhancer

HUDEP-2 cells with stable Cas9 expression were transduced at low multiplicity with the human sgRNA library lentivirus pool while in expansion medium. Control transductions were performed to ensure transduction rate did not exceed 50%. Cell numbers were maintained throughout the experiment at levels adequate to exceed 1000X representation of the library. 10 µg/mL blasticidin (Sigma) and 1 µg/mL puromycin (Sigma) were added 24 hours after transduction to select for lentiviral library integrants in cells with Cas9. Cells were cultured in expansion media for one week followed by differentiation media for an additional week.

Intracellular staining was performed by fixing cells with 0.05% glutaraldehyde (grade II) (Sigma) for 10 minutes at room temperature. Cells were centrifuged for 5 minutes at 600 g and

then resuspended in 0.1% Triton-X 100 (Life Technologies) for 5 minutes at room temperature for permeabilization. Triton X-100 was diluted with phosphate buffered saline (PBS) with 0.1% BSA and then centrifuged at 600 g for 15 minutes. Cells were stained with anti-human antibodies for HbF (clone HbF-1 with FITC or APC conjugation; Life Technologies) and β-hemoglobin antibody (clone 37-8 with PerCP-Cy5 or PE conjugation; Santa Cruz) for 20 minutes in the dark. Cells were washed to remove unbound antibody prior to FACS analysis. 0.2 µg HbF and 2 µg of HbA (β-hemoglobin) antibodies were used per 5 million cells. Control cells exposed to a nontargeting sgRNA sample and *BCL11A* exon 2 were used as negative and positive controls respectively to establish flow cytometry conditions. Populations of cells with the top and bottom 10% of expression of HbF were sorted by FACS.

After sorting the HbF-high and HbF-low pools, library preparation and deep sequencing was performed as previously described[84]. Briefly, genomic DNA was extracted using the Qiagen Blood and Tissue kit. Herculase PCR reaction (Agilent) using lentiGuide-Puro specific primers (5'-AATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCG-3' and 5'-CTTTAGTTTGTATGTCTGTTGCTATTATGTCTACTATTCTTTCCC-3') including a handle sequence was performed as follows: Herculase II reaction buffer (1x), forward and reverse primers (0.5 µM each), dimethyl sulfoxide (DMSO) (8%), deoxynucleotide triphosphates (dNTPs) (0.25 mM each), Herculase II Fusion DNA Polymerase (0.5 reactions) using the following cycling conditions: 95ºC for 2 minutes; 20 cycles of 95ºC for 15 seconds, 60ºC for 20 seconds, 72ºC for 30 seconds; 72ºC for 5 minutes. Multiple reactions of no more than 200 ng each were used to amplify from 6.6 ug gDNA (~1e6 cell genomes) per pool. Samples were subjected to a second PCR using handle-specific primers[84] to add adaptors and indexes to each sample using the following conditions: Herculase II reaction buffer (1x), forward and reverse primers (0.5 µM each), dNTPs (0.25 mM each), Herculase II Fusion DNA Polymerase (0.5 reactions) with the following cycling conditions: 95ºC for 2 minutes; 25 cycles of 95ºC for 15 seconds, 60ºC for 20 seconds, 72ºC for 30 seconds; 72ºC for 5 minutes. PCR products were

run on an agarose gel and the band of expected size was gel purified. Illumina MiSeq 150 bp paired end sequencing was performed.

sgRNA sequences present in the plasmid pool as well as in the HbF-high and HbF-low pools were enumerated. Guide sequences were mapped to the guides comprising the sgRNA library without allowing mismatches. Total reads were normalized to library sequencing depth. Cellular dropout score was determined by calculating (1) the ratio of normalized reads in the cells at end of experiment (average of reads in the HbF-high and HbF-low pools) to reads in the plasmid pool; (2) $\log_2$ transformation; and (3) median of biological replicates. HbF enrichment score was determined by calculating (1) the ratio of normalized reads in the HbF-high compared to reads in the HbF-low pools; (2) $\log_2$ transformation; and (3) median of biological replicates. After exclusion of sgRNAs with dropout scores < 2^{-3} and NAG PAM sgRNAs, a Q-Q plot was made with a line fitted through the first and third quantiles using R software. HbF enrichment scores and cellular dropout scores were compared by Spearman rank correlation. sgRNA sequences were mapped to the human genome (hg19) with cleavage positions set to between positions 17 and 18 given PAM positions 21-23. For visual comparisons to targeting sgRNAs, nontargeting sgRNAs were pseudomapped each separated by 5 bp.

Validation in primary human CD34$^+$ hematopoietic stem and progenitor cells (HSPCs)

Primary human CD34$^+$ HSPCs from G-CSF mobilized healthy adult donors were obtained from the Center of Excellence in Molecular Hematology at the Fred Hutchinson Cancer Research Center, Seattle, Washington. CD34$^+$ HSPCs were subject to erythroid differentiation liquid culture as previously described[125]. Briefly, HSPCs were thawed on day 0 into erythroid differentiation medium (EDM) consisting of IMDM supplemented with 330 µg/mL holo-human transferrin, 10 µg/mL recombinant human insulin, 2 IU/mL heparin, 5% human solvent detergent pooled plasma AB, 3 IU/mL erythropoietin, 1% L-glutamine, and 2% penicillin/streptomycin. During days 0-7 of culture, EDM was further supplemented with 10^{-6} M hydrocortisone (Sigma),

100 ng/mL human SCF, and human IL-3 (R&D). During days 7-11 of culture, EDM was supplemented with 100 ng/mL human SCF only. During days 11-18 of culture, EDM had no additional supplements.

HSPCs were transduced with lentiCas9-Blast (Addgene plasmid ID 52962) 24 hours after thawing in the presence of 10 μM 16,16-dimethylprostaglandin E2 (PGE2; Cayman Chemical). At 48 hours after thawing, medium was changed and cells were transduced with lentiGuide-Puro or lentiGuide-Crimson cloned with relevant sgRNA sequence in the presence of 10 μM PGE2. Three independent transductions were performed per sgRNA. At 72 hours after thawing, medium was changed and HSPCs were selected with 10 μg/mL blasticidin and 1 μg/mL puromycin or 10 μg/mL blasticidin followed by sorting for lentiGuide-Crimson^{+} cells on day 16 of culture. Blasticidin and/or puromycin selection occurred from days 3 to 8 of culture.

Differentiation was assessed on day 18 of culture using anti-human antibodies against the transferrin receptor (CD71) [Clone OKT9 with FITC conjugation; eBioscience] and glycophorin A (CD235a) [Clone HIR2 with PE conjugation; eBioscience]. Enucleation was assessed using 2 μg/mL of the cell-permeable DNA dye Hoescht 33342 (Life Technologies). CD235a^{+}Hoescht 33342^{-} cells were determined to be enucleated erythroid cells. Cells were intracellularly stained for HbF and HbA on day 18 of culture as described above. 50,000-100,000 cells were centrifuged onto microscope slides at 350rpm for 4 minutes. Slides were stained with Harleco May-Grünwald stain (Millipore) for two minutes, Giemsa stain (Sigma) for 12 minutes, and two water washes for 30 seconds each. Slides were air dried and then cover-slipped using Fisher Chemical Permount Mounting Medium (Fisher). RNA isolation and RT-qPCR was performed as above. Gene expression was normalized to that of GAPDH. All gene expression data represents the mean of at least three technical replicates.

PCR primers were designed to amplify the genomic cleavage site for a given sgRNA. Resulting PCR products were subjected to Sanger sequencing. Sequencing traces were used for editing quantification using a previously described publically available tool[126].

Computational analysis

Human erythroid H3K27ac ChIP-seq was obtained from Xu et al[104] and mouse erythroid H3K27ac ChIP-seq was obtained from Kowalczyk et al[127] and Dogan et al[128]. We uniformly processed all the datasets using the same pipeline with the same criteria to call super-enhancers. Specifically, we started from raw reads and realigned each dataset with Bowtie2 with the default parameters. We then removed duplicate reads with the Picard Suite. To call the peaks we used MACS2 in the narrow mode. Finally to call the super-enhancers we used the ROSE algorithm with the default parameters[107]. Using these settings, peaks closer than 12.5 kb are stitched together and then ranked based on the H3K27ac intensity. To assign super-enhancers to genes we used again ROSE with default settings. In particular, the tool reports three categories of genes for each super-enhancer: 1) overlapping genes - genes for which the gene body region overlaps a super-enhancer; 2) proximal genes - genes close to a SE considering a window of 50kb; 3) closest gene - closest gene considering its TSS and the center of the super-enhancer. To generate a Venn diagram of genes for super-enhancer datasets, we used the union of these three gene categories.

Hidden Markov Model (HMM) segmentation was performed to automatically segment the enrichment score signals into enhancer regions with Active, Repressive and Neutral effect. We designed a HMM with 3 states using the GHMM package (http://ghmm.sourceforge.net/). To learn the HMM parameters we used the Baum-Welch algorithm. To find the best segmentation for each region we used the Viterbi algorithm. The emission probability for each state was modeled as a Gaussian distribution and all the possible transitions between states were allowed as shown in Supplemental Figure 3.3a. Since the signal was not obtained with a constant genomic resolution, we interpolated and smoothed the signal using a Gaussian kernel over 12 bp and applied the HMM to the smoothed signal. To set the initial parameters, we used the 1%, 50% and 99% percentile of the smoothed signal for the prior of the means of the Repressive,

Neutral and Active states respectively, while the prior for the standard deviation was set to 0.001 for all the three states.

Motif analysis was performed to evaluate the human and mouse enhancer regions for potential binding sites for known transcription factors (TFs). We used the FIMO software[129] with a *P*-value threshold of $< 10^{-4}$. For each region we extracted sequences using the hg19 and mm9 assemblies respectively for human and mouse. The motif database was the latest version of the JASPAR database[130].

Deep sequencing paired-end reads of genomic amplicons from genome editing target sites were first filtered for reads with PHRED quality score <30, merged with the FLASH (Fast Length Adjustment of SHort reads) software, and subsequently aligned to a reference amplicon using the *needle* aligner from the EMBOSS suite (http://emboss.sourceforge.net/) to quantify insertions and deletions. Per nucleotide frequency of deletion of a position, insertion directly adjacent to the position, or no mutation at the position was quantitated using CRISPResso (https://github.com/lucapinello/CRISPResso).

Pooled CRISPR-Cas9 screen for high resolution functional mapping of mouse *Bcl11a* enhancer

Murine erythroleukemia (MEL) cells were cultured in DMEM supplemented with 10% FBS, 1% L-glutamine, and 2% penicillin-streptomycin as previously described[67]. Cell lines tested negative for mycoplasma contamination. εy:mCherry reporter MEL cells with stable Cas9 expression were transduced at low multiplicity with the mouse sgRNA library lentivirus pool (Supplemental Data). Control transductions were performed to ensure transduction rate did not exceed 50%. Cell numbers were maintained throughout the experiment at levels adequate to exceed 1000X representation of the library. 10 μg/mL blasticidin and 1 μg/mL puromycin were added 24 hours after transduction to select for lentiviral library integrants in cells with Cas9. Subsequently cells were cultured for two weeks. The top and bottom 5% of εy-mCherry-expressing cells exposed to

the library were sorted by FACS. A nontargeting sgRNA sample was used as a negative control and *Bcl11a* exon 2 as a positive control to establish flow cytometry conditions. After sorting, library preparation and deep sequencing were performed as described for the human library[84].

sgRNA sequences present in the Hbb-εy:mCherry-high and Hbb-εy:mCherry-low pools were enumerated. Cellular dropout and εy enrichment scores were calculated analogously to the human screen. sgRNA sequences were then mapped to the mouse genome (mm9).

Generation of genomic deletions in MEL cells

Deletions in MEL cells were generated using two sgRNA as previously described[60]. Briefly, sgRNA sequences were cloned into pX330 (Addgene plasmid ID 42230) using a Golden Gate assembly cloning strategy (Supplemental Data Table 3.1 and 3.4). MEL cells were electroporated with 5 μg of each pX330-sgRNA plasmid and 0.5 μg pmax-GFP (Lonza) in BTX electroporation buffer using a BTX electroporator (Harvard Apparatus). Approximately 48 hours post-electroporation, the top 1-3% of GFP^+ cells were sorted and plated clonally at limiting dilution. Clones were allowed to grow for 7-10 days. Clones were screened for deletion by conventional PCR using the same strategy as with the HUDEP-2 cells (Supplemental Table 3.2). Inversion clones were identified by PCR as previously described[60] (Supplemental Table 3.3).

Generation of genomic deletions in mouse embryonic stem cells (mESCs)

mESCs were maintained on irradiated mouse embryonic fibroblasts (GlobalStem) and cultured in high glucose DMEM supplemented with 20% FBS, L-glutamine, penicillin/streptomycin (Life Technologies), non-essential amino acids (Life Technologies), nucleosides, β-mercaptoethanol (Sigma), and leukemia inhibitory factor (Millipore). Cells were passaged using 0.25% trypsin (Life Technologies).

The *Bcl11a* +62 deletion mice were derived from CRISPR-Cas9 modified CJ9 ES cells. Using Amaxa ES Cell transfection reagent (Lonza), two million mESCs cells were electroporated with 2 μg of each pX330 plasmid vector containing individual target sequences flanking the +62 site along with 0.5 μg of a GFP plasmid. After 48 hours, the top 5% of GFP expressing cells were sorted, plated on irradiated fibroblasts and maintained. Individual ES cell colonies were then picked and screened for biallelic deletion using the same strategy as HUDEP-2 and MEL cells[60]. DNA for screening CRISPR-Cas9 modified clones was obtained from gelatin adapted ES cell clones to avoid genomic contamination from the fibroblasts. Correctly targeted clones with greater than 80% normal karyotype were used to generate mice. Clones were injected into embryonic day 3.5 (E3.5) C57Bl6 blastocysts and implanted into pseudo-pregnant females.

The β-YAC mouse line (A20), previously described as containing a transgene encompassing ~150 kb of the human β-globin locus[131], was used to analyze human globin expression. The mouse line was maintained in a hemizygous state and bred with *Bcl11a* +62 deletion mice. Sufficient matings were established to ensure adequate homozygotes for analysis.

Mouse cell and tissue analysis

For developmental hematopoiesis, fetal liver cells were taken at E12.5, E14.5, E16.5, and E18.5 and mechanically dissociated to form single cell suspensions from which RNA was extracted using the RNeasy Plus Mini Kit (Qiagen) and analyzed. At E16.5, fetal liver were also stained with CD19-PerCP-Cy5.5 (Clone 1D3; eBioscience), B220-APC (RA3-6B2; Biolegend), CD71-PE (Clone C2; BD Biosciences), and Ter119-FITC (Clone Ter119; BD Biosciences) to isolate B cells ($B220^{+}CD19^{+}$) and erythroid cells ($Ter119^{+}CD71^{+}$) by FACS for RNA extraction and BCL11A quantification. Additionally, flow cytometry was used to analyze fetal liver from E18.5 embryos. Single cell suspensions were stained with IgM-FITC (Clone II-41; eBioscience), CD19-

PerCP-Cy5.5, (Clone 1D3; eBioscience), CD43-PE (Clone S7; eBioscience), AA4.1-PE-Cy7 (Clone AA4.1; BD Biosciences), B220-APC, (RA3-6B2; Biolegend), and DAPI (Invitrogen). For adult hematopoietic assays, peripheral blood was obtained from the tail vein of four week old mice. Blood was collected in EDTA-coated tubes, red cells removed by 2% dextran (Sigma), residual red cells lysed with ammonium chloride solution (Stem Cell Technologies) and stained with the following anti-mouse antibodies: CD3e-FITC (Clone 145-2C11; Biolegend), CD19-PerCP-Cy5.5 (Clone 1D3; eBioscience), CD71-PE (Clone C2; BD Biosciences), NK1.1-PE-Cy5 (Clone PK136; Biolegend), Ter119-APC (Clone TER-119; Biolegend), Gr-1-eF450 (Clone RB6-8C5; eBioscience), B220-BV605 (RA3-6B2; Biolegend), Mac-1-BV510 (Clone M1/70; Biolegend), and 7-AAD (BD Biosciences). Fetal brain analysis was conducted on whole brains from E16.5 mouse embryos on ice cold PBS. Tissue was directly lysed into the RLT plus buffer (Qiagen) and total RNA extracted according to manufacturer's instructions provided in the RNeasy Plus Mini Kit. RT-qPCR performed as above, with gene expression normalized to Gapdh. All gene expression data represents the mean of at least three technical replicates. All animal experiments were conducted under the approval of the local Institutional Animal Care and Use Committee.

Cloning lentiCas9-Venus

Venus template[132] was PCR amplified to add BamHI-HF (5') and EcoRI-HF (3') restriction sites for cloning purposes using the following conditions: KOD buffer (1x), $MgSO_4$ (1.5 mM), dNTPs (0.2 mM each), forward primer (0.3 μM; GGCCGGCCggatccGGCGCAACAAACTTCTCTCTGCTGAAACAAGCCGGAGATGTCGAAGAGAATCCTGGACCGATGGTGAGCAAGGGCGAGGA), reverse primer (0.3 μM; GGCCGGCCgaattcTTACTTGTACAGCTCGTCCA), and KOD Hot Start DNA Polymerase (0.02 U/μL) (Millipore). KOD PCR reaction used the following cycling conditions: 95°C for 2 minutes; 50 cycles of 95°C for 20 seconds, 60°C for 20 seconds, and 70°C for 30 seconds; 60°C for 5

minutes. PCR products were purified (QIAquick PCR Purification Kit, Qiagen) and blunt end cloned with Zero Blunt PCR cloning kit (Invitrogen). PCR-blunt cloned products and lentiCas9-Blast (Addgene plasmid ID 52962) were separately digested with BamHI-HF (New England Biolabs) and EcoRI-HF (New England Biolabs) in 1x Buffer CutSmart at 37°C (New England Biolabs). Digest of lentiCas9-Blast was performed to remove the blasticidin cassette. Then digested PCR product was ligated into the lentiCas9 backbone.

Cloning lentiGuide-Crimson

E2-Crimson template (Clontech) was PCR amplified to add BsiWI (5') and MluI (3') restriction sites for cloning purposes using the following conditions: KOD buffer (1x), $MgSO_4$ (1.5 mM), dNTPs (0.2 mM each), forward primer (0.3 μM; GGCCGGCCCGTACGcgtacgGCCACCATGGATAGCACTGAGAACGTCATCAAGCCCTT), reverse primer (0.3 μM; GGCCGGCCacgcgtCTACTGGAACAGGTGGTGGCGGGCCT), and KOD Hot Start DNA Polymerase (0.02 U/μL). KOD PCR reaction used the following cycling conditions: 95°C for 2 minutes; 50 cycles of 95°C for 20 seconds, 60°C for 20 seconds, and 70°C for 30 seconds; 60°C for 5 minutes. PCR products were purified (QIAquick PCR Purification Kit) and cloned with Zero Blunt PCR cloning kit. Cloned products and lentiGuide-puro were separately digested with BsiWI (New England Biolabs) and MluI (New England Biolabs) in 1x Buffer 3.1 at 37°C (New England Biolabs). Digest of lentiGuide-Puro (Addgene plasmid ID 52963) was performed to remove the puromycin cassette. Then digested PCR product was ligated into the lentiGuide backbone.

Cloning sgRNAs

lentiGuide-Puro (Addgene plasmid ID 52963) was digested with BsmBI in 1X Buffer 3.1 at 37°C (New England Biolabs) for linearization. One unit of TSAP thermosensitive Alkaline Phosphatase (Promega) was added for 1 hour at 37°C to dephosphorylate the linearized

lentiGuide and then TSAP was heat inactivated at 74°C for 15 minutes. Linearized and dephosphorylated lentiGuide was run on an agarose gel and gel purified. sgRNA-specifying oligos were phosphorylated and annealed using the following conditions: sgRNA sequence oligo (10 µM); sgRNA sequence reverse complement oligo (10 µM); T4 ligation buffer (1x) (New England Biolabs); and T4 polynucleotide kinase (5 units) (New England Biolabs) with the following temperature conditions: 37 °C for 30 min; 95 °C for 5 min; and then ramp down to 25 °C at 5 °C/min. Annealed oligos were ligated into lentiGuide in a 1:3 ratio (vector:insert) using T4 ligation buffer (1X) and T4 DNA Ligase (750 Units) (New England Biolabs. Plasmids were verified by sequencing using a U6F promoter forward primer CGTAACTTGAAAGTATTTCGATTTCTTGGC.

sgRNA-specifying oligos using sgRNA sequences from the screen library (Supplemental Data) were obtained and cloned as described into either lentiGuide-Puro or lentiGuide-Crimson. sgRNA constructs were used to produce lentivirus and transduce HUDEP-2 with stable Cas9 expression. Bulk cultures were incubated for 7-10 days with 10 µg/mL blasticidin and 1 µg/mL puromycin selection to allow for editing. Then bulk cultures were plated clonally at limiting dilution. Clones were allowed to grow for approximately 14 days and then genomic DNA was extracted using 50 µL QuickExtract DNA Extraction Solution per well.

lentiTandemGuide cloning

lentiGuide-sgRNA1 was digested with PspXI and XmaI at 37°C for four hours (New England Biolabs). Digests were run on an agarose gel and gel purified. lentiGuide-sgRNA2 was linearized using NotI (New England Biolabs). The hU6 promoter and sgRNA chimeric backbone for lentiGuide-sgRNA2 was PCR amplified using the following conditions: KOD buffer (1x), $MgSO_4$ (1.5 mM), dNTPs (0.2 mM each), forward primer (0.3 µM; GGCCGGCCgctcgaggGAGGGCCTATTTCC), reverse primer (0.3 µM; CCGGCCGGcccgggTTGTGGATGAATACTGCCATTT), and KOD Hot Start DNA Polymerase

(0.02 U/μL) (Millipore). KOD PCR reaction used the following cycling conditions: 95°C for 2 minutes; 50 cycles of 95°C for 20 seconds, 60°C for 20 seconds, and 70°C for 30 seconds; 60°C for 5 minutes. PCR products were purified (QIAquick PCR Purification Kit), blunt ended cloned with Zero Blunt PCR cloning kit, transformed, and plated. Colonies were screened by digesting minipreps with EcoRI. Mini-preps were then digested with PspXI and XmaI as described above followed by PCR purification. Following PCR purification, sgRNA2 was ligated into digested lentiGuide-sgRNA1. Sequence verified with following primers: GGAGGCTTGGTAGGTTTAAGAA and CCAATTCCCACTCCTTTCAA.

Generation of HUDEP-2 with stable Cas9

lentiCas9-Blast (Addgene plasmid ID 52962) or lentiCas9-Venus were produced as described above and used to transduce HUDEP-2 cells. Transduced cells were selected with 10 μg/mL blasticidin or Venus$^+$ cells were sorted. Functional Cas9 was confirmed using the pXPR-011 (Addgene plasmid ID 59702) GFP reporter assay as previously described[133].

Generation of εy:mCherry reporter MEL cells

A reporter MEL line in which mCherry was been knocked into the *Hbb-y* locus was created (Supplemental Figure 3.5a). Briefly, a TALEN-induced DSB was created adjacent to the *Hbb-y* transcriptional start site. A targeting vector with mCherry and a neomycin cassette were introduced through homology directed repair. Homology arms included mm9 sequences from chr7:111,001,667-111,002,675 and chr7:111,000,661-111,001,666. Cre-mediated recombination was utilized to remove the neomycin cassette. Long-range PCR spanning each homology arm was utilized to ensure appropriate targeted integration. Cells were tested upon *Bcl11a* disruption by RT-qPCR and flow cytometry to confirm expected effects on εy:mCherry derepression. Subsequently CRISPR-Cas9 was used as described above to produce cells with

monoallelic composite enhancer deletion to maximize screening sensitivity for enhancer disruption.

Generation of MEL cells with stable Cas9 expression

lentiCas9-Blast (Addgene plasmid ID 52962) lentivirus were produced as described above and used to transduce MEL cells. Transduced cells were selected with 10 μg/mL blasticidin. Functional Cas9 was confirmed using the pXPR-011 (Addgene plasmid ID 59702) GFP reporter assay as previously described[133].

RESULTS

Human composite enhancer

Recently we observed that common genetic variants associated with HbF ($\alpha_2\gamma_2$) level and β-hemoglobin disorder clinical severity mark an adult developmental stage- and erythroid-lineage specific intronic enhancer of *BCL11A*[67], a validated repressor of HbF and therapeutic target for β-hemoglobin disorders[67,134–136]. This composite human enhancer is composed of three DNase I hypersensitive sites (DHSs), termed h+55, h+58, and h+62, based on distance in kilobases from the transcriptional start site (TSS)[67]. The most highly trait-associated haplotype is defined by two SNPs, rs1427407 within h+62 and rs7606173 within h+55 (Supplemental Figure 3.1a). Previously we showed that this enhancer possessed ectopic erythroid-restricted, adult-stage-specific enhancer activity[67]. Moreover, the mouse ortholog of the composite enhancer, defined by primary sequence homology, shared erythroid enhancer chromatin signature, and syntenic position relative to coding sequences, was shown to be required for BCL11A expression and embryonic globin gene repression in a mouse erythroid cell line but dispensable in a mouse B-lymphoid cell line[67].

To evaluate the requirement for human *BCL11A* enhancer sequences, we utilized HUDEP-2 cells, an immortalized human $CD34^+$ hematopoietic stem and progenitor cell (HSPC)-

derived erythroid precursor cell line that expresses BCL11A and predominantly β- rather than γ-globin[122]. We used the clustered regularly interspaced short palindromic repeat (CRISPR)-Cas9 nuclease system to generate clones of HUDEP-2 cells with deletion of the 12-kb *BCL11A* composite enhancer by introduction of a pair of chimeric single guide RNAs (sgRNAs). Enhancer deletion resulted in near complete loss of BCL11A expression and induction of γ-globin and HbF protein to similar levels as cells with *BCL11A* knockout (Figure 3.1a-c), consistent with the possibility that these sequences could serve as targets for therapeutic genome editing for HbF reinduction for the β-hemoglobinopathies[137]. Although targeted deletions by paired double strand breaks (DSBs) may be achieved by genome editing, competing genomic outcomes include local insertion/deletion (indel) production at each cleavage site as well as inversion of the intervening segment[47,48,60,68,138].

Tiled pooled enhancer editing *in situ*

We hypothesized that composite enhancers may be composed of a functional hierarchy with essential and dispensable constituent components. A functional hierarchy might enable enhancer disruption by a single DSB at a critical region followed by nonhomologous end joining (NHEJ) repair with indels. In fact, the hypothesis that a prevalent mechanism of trait associations is enhancer variation rests on the premise that single nucleotide changes themselves may substantively modulate enhancer function. Therefore we reasoned that a tiling set of sgRNAs could uncover critical enhancer regions by disruption of nearly all sequences within an enhancer based on the typical outcome of Cas9 cleavage and NHEJ repair, an indel spectrum with frequent deletions of up to 10 bp from the cleavage position[47,48,60,68,93].

We designed all possible sgRNAs within the human *BCL11A* composite enhancer DHSs (Figure 3.1d, e) as restricted only by the presence of the SpCas9 NGG protospacer adjacent motif (PAM), which restricts cleavage at an average 1/8 frequency at each genomic position[47,93]. The NGG PAM restricted sgRNAs had a median gap between adjacent genomic cleavages of 4

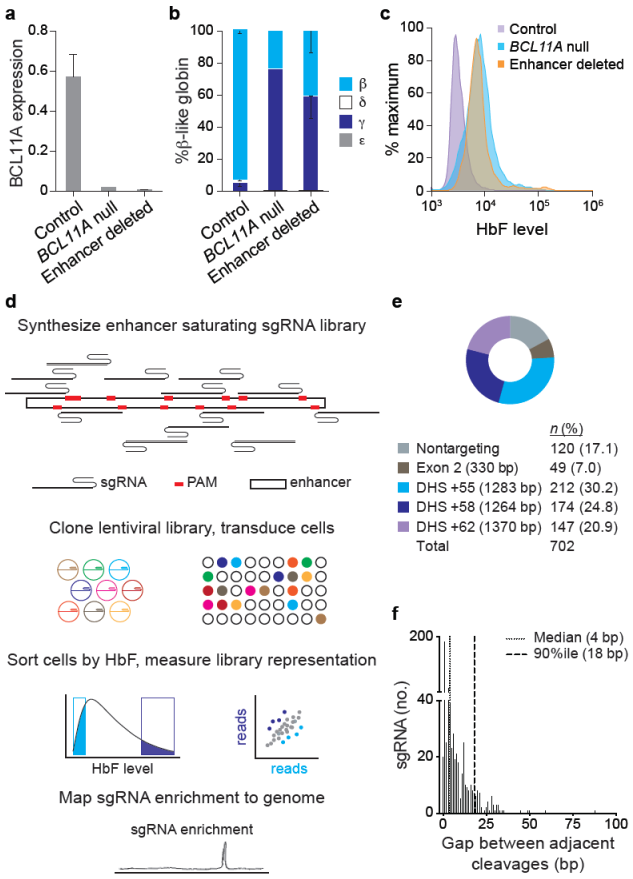


Figure 3.1: Tiled pooled in situ CRISPR-Cas9 BCL11A enhancer screen. **a-c**, Deletion of the human composite BCL11A enhancer in HUDEP-2 cells demonstrates its necessity for BCL11A expression (normalized to GAPDH), repression of γ-globin mRNA, and repression of HbF. Error bars show s.e.m. **d**, Workflow of CRISPR-Cas9 enhancer screen showing library synthesis, delivery, and analysis. **e**, Human NGG PAM sgRNA library distribution. **f**, Gaps between adjacent genomic cleavages for NGG PAM sgRNAs targeting BCL11A exon-2, h+55, h+58, and h+62.

bp and 90th percentile of 18 bp (Figure 3.1f), which suggested that this strategy could approach saturation mutagenesis *in situ*. We included nontargeting sgRNAs as negative controls as well as sgRNAs tiling exon-2 of *BCL11A* as positive controls (Figure 3.1e). The library was successfully cloned to a lentiviral vector. The basic experimental schema was to transduce HUDEP-2 cells with the lentiviral library at low multiplicity such that nearly all selected cells contained a single integrant (Figure 3.1d). Following expansion, differentiation, sorting by HbF level, genomic DNA isolation, and deep sequencing of integrated sgRNAs, an HbF enrichment score was calculated for each sgRNA by comparing its representation in HbF-high and HbF-low pools (see Supplementary Information and Supplemental Figure 3.2 for additional technical details).

We mapped the HbF enrichment score of each sgRNA to its predicted position of genomic cleavage (Figure 3.2a). The majority of enhancer targeting sgRNAs showed no significant enrichment or depletion from the HbF-high pool. The enriching sgRNAs colocalized to discrete genomic positions. For example, we observed a cluster of sgRNAs at h+62 with modest enrichment, a cluster at h+55 with moderate enrichment (as well as adjacent clusters with depletion), and a cluster at h+58 with marked enrichment. Of note, we observed 10

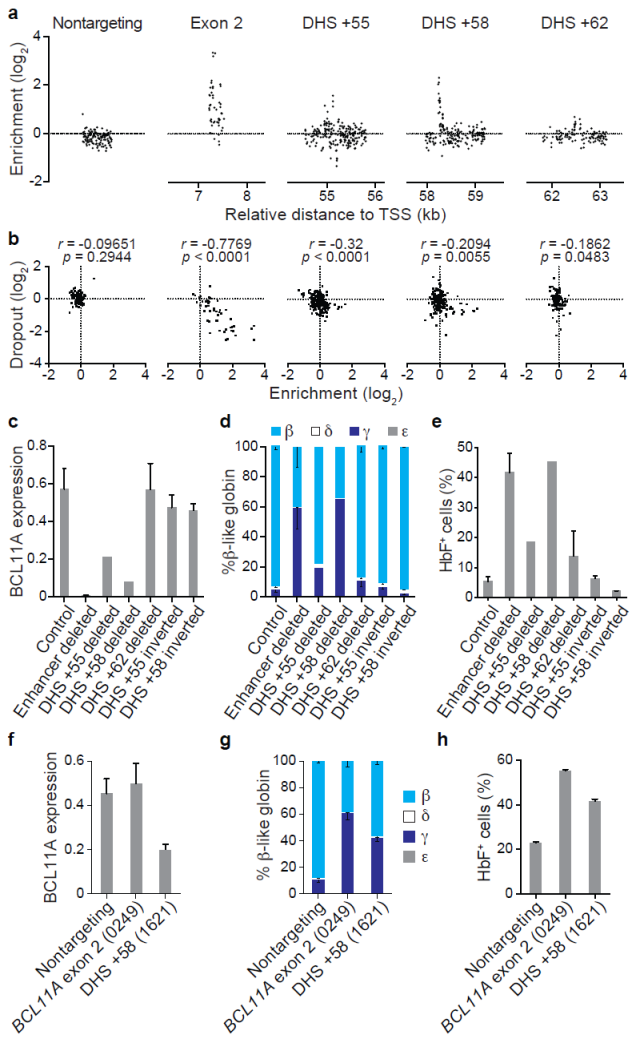


Figure 3.2: Functional mapping of the *BCL11A* enhancer. a, Mapping sgRNA HbF enrichment scores relative to genomic cleavage positions. Nontargeting sgRNAs pseudo-mapped with 5 bp spacing. **b**, Correlation between cellular dropout and HbF enrichment scores. **c-e**, BCL11A expression normalized to GAPDH, β-like globin expression, and HbF$^+$ fraction in HUDEP-2 cells with deletion or inversion of individual DHSs. **f-h**, BCL11A expression normalized to GAPDH, β-like globin expression, and HbF$^+$ fraction in primary human erythroid precursors transduced with Cas9 and individual sgRNAs. Error bars represent s.e.m. (**c**, **d**, **f**, **g**) or s.d. (**e**, **h**).

sgRNAs at h+58 with cleavage positions within 42 bp each with HbF enrichment scores exceeding 0.99, the median enrichment score of *BCL11A* exon-2 targeting sgRNAs.

Exon-2 targeted sgRNAs showed a linear correlation between HbF enrichment and cellular dropout, suggesting sgRNAs that result in complete knockout of *BCL11A* lead to a reduced rate of cell accumulation inseparable from HbF derepression (Figure 3.2b). In contrast, the sgRNAs at h+58 associated with marked HbF enrichment showed blunted impact on dropout (Figure 3.2b). This finding could be consistent with a low residual level of BCL11A adequate to promote cellular accumulation but inadequate to suppress HbF.

To corroborate these findings, we introduced two sgRNAs to the HUDEP-2/Cas9 cells to produce targeted deletion or inversion clones[60]. Deletion of h+58 phenocopied deletion of the composite enhancer and deletion of h+55 had a moderate effect (while deletion of h+62 showed a nonsignificant trend towards a modest effect), consistent with the magnitude of top-scoring and colocalizing sgRNAs from the screen (Figure 3.2a, c-e). Inversion of the h+58 or h+55 sites had no significant effect on gene

expression, demonstrating that the *BCL11A* enhancer functions in an orientation-independent manner *in situ*, consistent with the classic enhancer definition[98] (Figure 3.2c-e).

To validate the findings from the HUDEP-2 cells, the top-scoring enhancer targeting sgRNA from the screen (#1621 at h+58) was tested in primary human erythroblasts by lentiviral transduction of human $CD34^+$ HSPCs exposed to *ex vivo* erythroid culture conditions. Consistent with the screen results, sgRNA-1621 resulted in downregulation of BCL11A expression and corresponding upregulation of γ-globin expression and increase in HbF^+ cells (Figure 3.2f-h). Notably, sgRNA-1621 did not alter surface marker profile, enucleation frequency, or cellular morphology (Supplemental Figure 3.3c). Together these results suggest proof-of-principle of an individual sgRNA targeting a noncoding element for therapeutic genome editing of β-hemoglobin disorders.

Primate-specific enhancer sequences

We applied a hidden Markov model (HMM) to the sgRNA enrichment score data to infer functionally important sequences within each DHS (Supplemental Figure 3.4a). This model defined three functional states, Active, Repressive, and Neutral, based on likelihood to encompass sequences that positively, negatively, and neutrally regulate target gene expression, respectively. The model identified functional states within each DHS (Figure 3.3a-c). At each of the three DHSs, the Active states were precisely located at regions with the highest degree of DNase I sensitivity.

The overall sequence conservation at the h+58 Active region appears both less intense and less distinct from flanking sequences as compared to those of h+62 and h+55 (Figure 3.3a-c). The top-scoring sgRNAs in the screen colocalize to 42 bp within h+58 (Figure 3.4, Supplemental Figure 3.5b). The third-highest scoring enhancer-targeted sgRNA (sgRNA-1617) mapped directly onto an apparent GATA1 motif, though below a genome-scale significance threshold ($P = 3.74 \times 10^{-4}$). The mouse orthologous sequence has a GATA1 motif *P*-value only

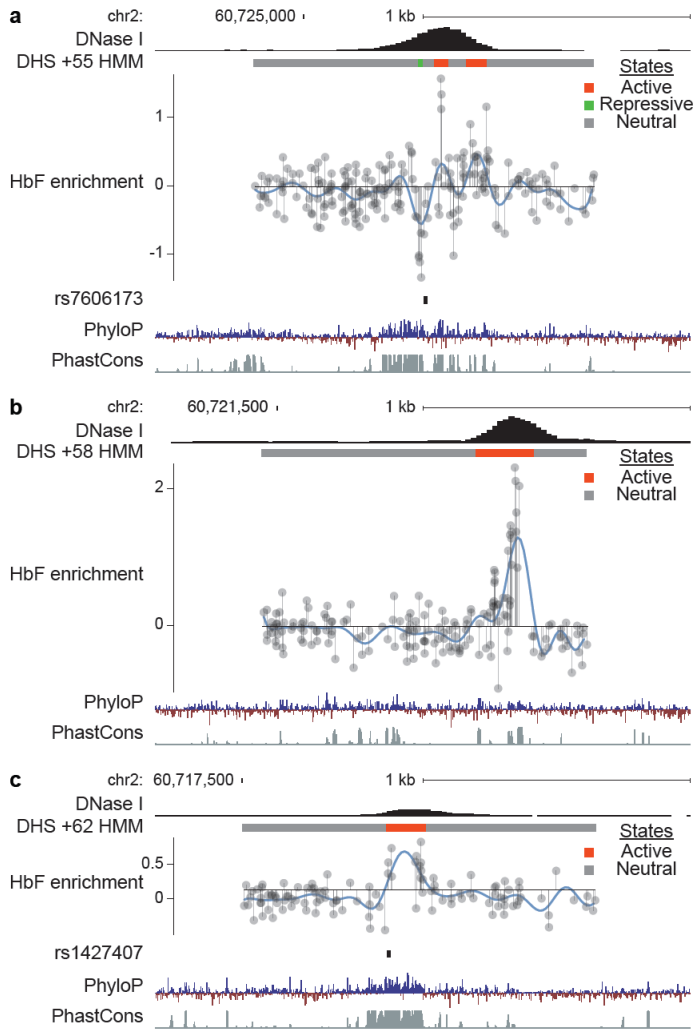


Figure 3.3: Inferred functional enhancer states relative to genomic features. a-c, Hidden Markov model segmentation of functional enhancer states. HbF enrichment scores shown throughout DHSs h+55, h+58, h+62 by gray lines and circles with blue line representing smoothed enrichment score. DNase I sequencing from primary human erythroblasts[67]. PhyloP (scale from -4.5 to 4.88) and PhastCons (from 0 to 1) estimates of evolutionary conservation among 100 vertebrates. Positions of SNPs rs7606173 and rs1427407 denoted which together define the haplotype most highly associated to HbF level[67].

modestly higher than the human ($p = 4.33 \times 10^{-4}$). This GATA1 motif appears to have relatively high vertebrate conservation, with exact human sequence identity in rabbits, pigs, dogs, and elephants. The top-scoring sgRNA (sgRNA-1621) mapped to a position 15 bp from this GATA1 motif (Figure 3.4). An additional four sgRNAs mapping between sgRNA-1621 and 1617 each had substantially elevated HbF enrichment scores. Underlying these sgRNAs were additional predicted motifs (i.e. RXRA, EHF, ELF1, and STAT1). Although these sequences showed a high level of conservation among primates, they showed high degeneracy among nonprimate vertebrates (Figure 3.4).

We tested the pattern of mutations observed upon treatment of cells with either sgRNA-1621 or sgRNA-1617 by deep sequencing. Each of these sgRNAs is sufficient to substantially induce HbF in human erythroid cells (Figure 3.2h; Supplemental Figure 3.3a, b). We sorted cells exposed to Cas9 and these sgRNAs into HbF-high and HbF-low pools. We determined the indel spectrum in each population by deep sequencing (Supplemental Figure 3.4b). As expected we observed indels clustering around the predicted cleavage positions. By comparing the per

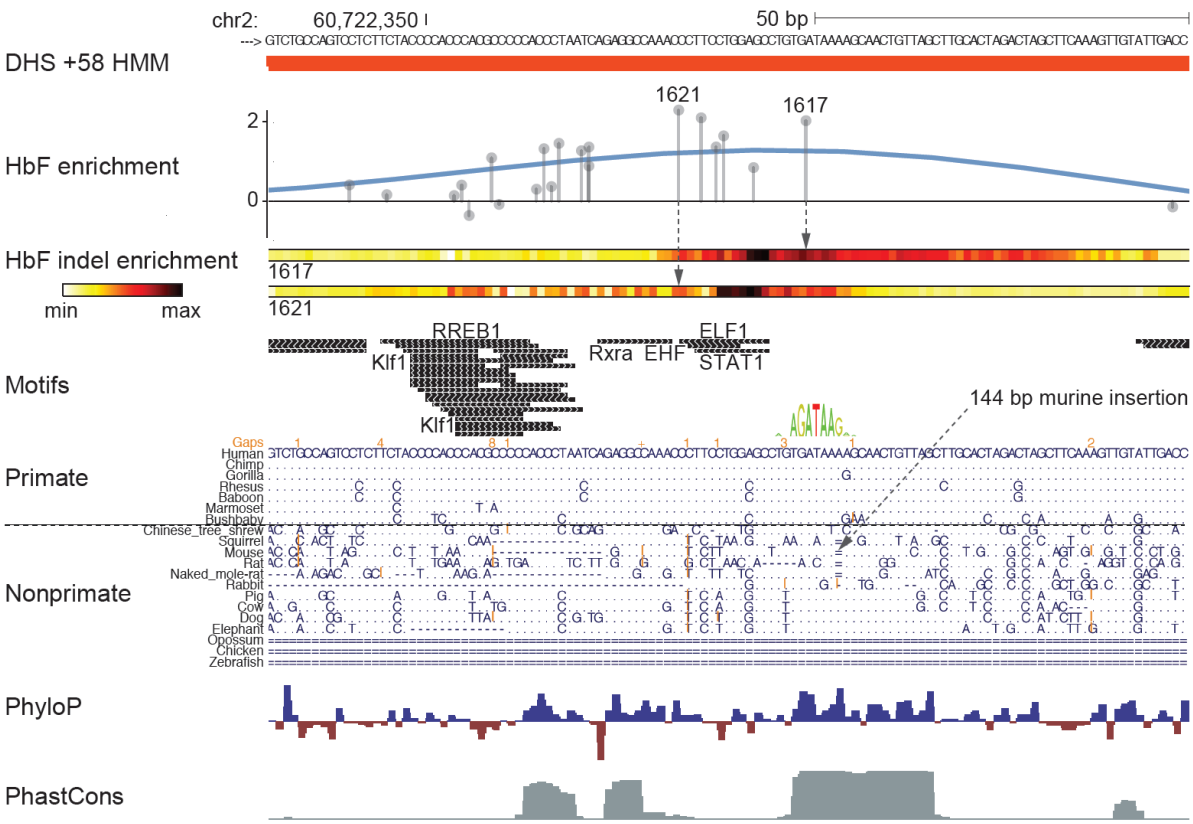


Figure 3.4: Primate-specific *BCL11A* enhancer functional core. DHS h+58 functional core defined by maximal HbF enrichment score and Active HMM state. HbF enrichment scores shown by gray lines and circles. HbF indel enrichment per nucleotide based on amplicon genomic sequencing of sorted cells exposed to either sgRNA-1617 or -1621. No common SNPs (MAF>1%) present at this region. JASPAR motifs ($P < 10^{-4}$) depicted in black with selected motifs annotated by TF based on known erythroid-specific function or genomic position. Gata1 motif LOGO at sgRNA-1617 cleavage position as described in text. Orthologous sequences listed from representative primates and nonprimates of distributed phylogeny. PhyloP (scale from -4.5 to 4.88) and PhastCons (from 0 to 1) estimates of evolutionary conservation among 100 vertebrates.

nucleotide indel ratio between cells from the HbF-high and HbF-low pools, we calculated a relative indel enrichment across the sequencing amplicon. Notably both sgRNAs yielded maximal HbF indel enrichment not precisely at the expected cleavage position but offset at shared intervening sequences (Figure 3.4). These sites of maximal HbF mutation enrichment mapped to 7 bp directly overlapping predicted motifs (Figure 3.4). Taken together, these data suggest that a conserved GATA1 motif scoring below the prediction threshold adjacent to primate-specific sequences form the core of an enhancer essential for human erythroid BCL11A expression and HbF repression.

Mouse enhancer dissection

To test functional conservation of the BCL11A enhancer, we examined the orthologous mouse *Bcl11a* enhancer in greater detail. Erythroid DNase I sensitivity is only observed at those sequences homologous to h+55 and h+62 and not h+58 (Supplemental Figure 3.6a), consistent with the reduced sequence homology within the h+58 Active region (Figure 3.3a-c). We performed a pooled CRISPR enhancer saturating mutagenesis screen in MEL εy:mCherry reporter cells, similar to the human screen described above (Supplemental Figure 3.6, 3.7; Supplementary Information).

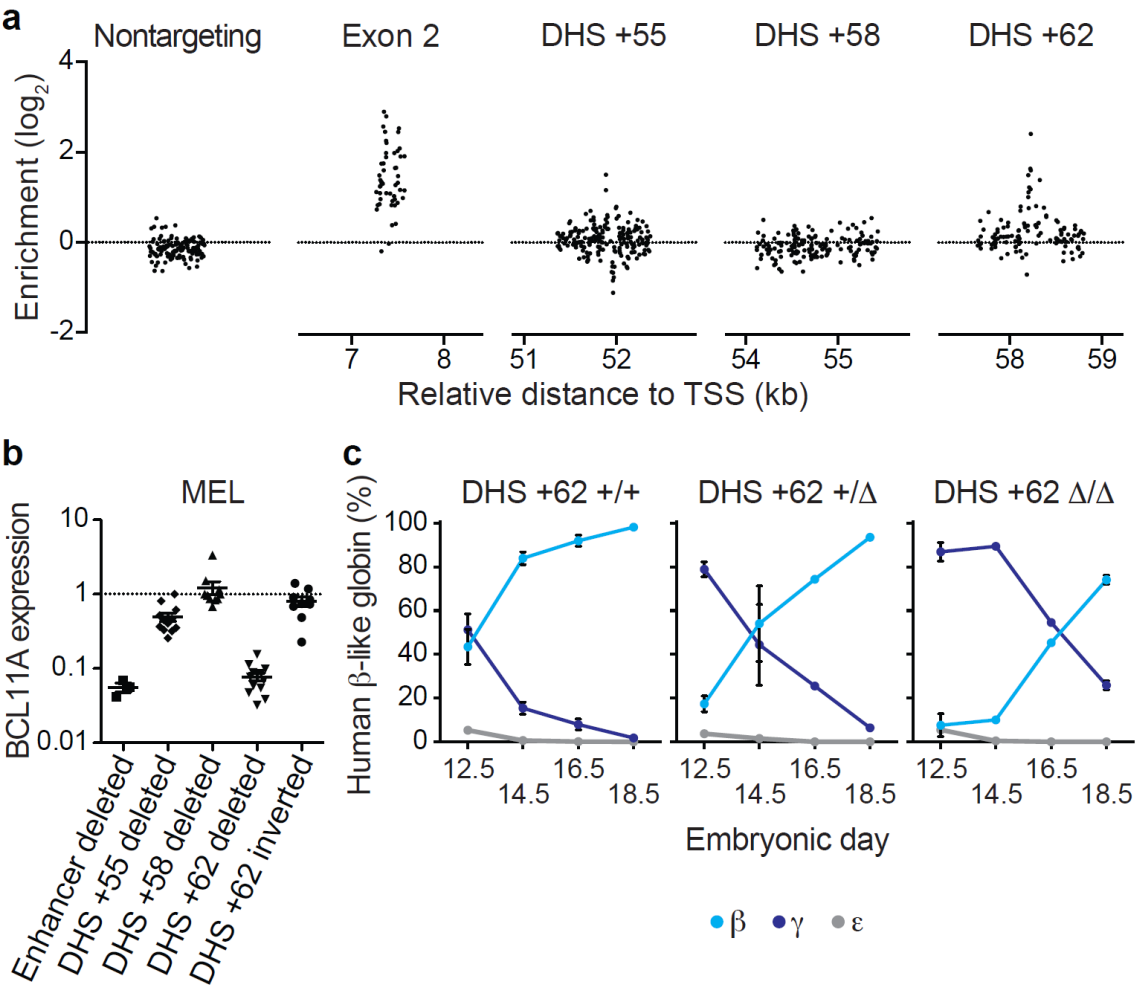


Figure 3.5: Functional sequence requirement at the mouse *Bcl11a* erythroid enhancer for *in vivo* hemoglobin switching. a, Mapping sgRNA εy enrichment scores to genomic cleavage positions. Nontargeting sgRNAs pseudo-mapped with 5 bp spacing. **b**, BCL11A expression in mouse erythroid clones with deletion or inversion of individual DHSs relative to nondeleted controls. **c**, Transgenic human β-like globin (each symbol represents the mean of at least 3 embryos) expression in β-YAC / +62 deletion mice. Error bars represent s.e.m.

Upon mapping the sgRNA cleavage positions to the genome, we again observed that the majority of enhancer targeting sgRNAs demonstrated no significant εy enrichment or depletion. We observed colocalization of sets of sgRNAs with εy enrichment (Figure 3.5a). There was a similar complex pattern at the m+55 ortholog as at h+55, with adjacent regions with enriching and depleting sgRNAs from the high-εy:mCherry pool at the DHS core. At the m+58 ortholog we did not observe any evidence of εy enriching or depleting sgRNAs. At the m+62 ortholog there was a marked peak, with five sgRNAs with εy enrichment scores exceeding 1.30, the median enrichment score of *Bcl11a* exon-2 targeting sgRNAs (Figure 3.5a). This potent impact of the m+62 ortholog was in contrast to the modest impact of individual sgRNAs or DHS deletion at h+62.

We used pairs of sgRNAs in the presence of Cas9 to produce MEL clones with deletions of various substituent elements at the *Bcl11a* enhancer (Figure 3.5b). Deletion of the DNase-insensitive m+58 ortholog had no apparent effect on BCL11A expression consistent with the pooled screen result. Deletion of the m+55 ortholog led to an approximately two-fold reduction in BCL11A expression (mean residual level 49%, $p<0.0001$), whereas deletion of the m+62 ortholog approached deletion of the entire composite enhancer in terms of reduction in BCL11A expression (mean residual levels of 8% ($p<0.0001$) and 6% ($p<0.0001$) respectively, Figure 3.5b; also see Supplementary Information and Supplemental Figure 3.8, 3.9). In addition, clones in which the m+62 ortholog was inverted showed no change in BCL11A expression, suggesting that the mouse, like the human, enhancer functions independent of orientation *in situ* (Figure 3.2c-e; 3.5b).

Erythroid-restricted function *in vivo*

To substantiate the importance of the m+62 ortholog for BCL11A expression as well as to validate BCL11A enhancer disruption as a therapeutic strategy, we generated mice deficient for the *Bcl11a* m+62 ortholog. We used the same Cas9 and paired sgRNA deletion strategy in

mouse embryonic stem cells (mESCs), from which we derived mice with germline transmission of the enhancer deletion. Previous studies have demonstrated essential roles for *Bcl11a* in structural development of the central nervous system and B-lymphopoiesis[139–141]. Strikingly, unlike conventional *Bcl11a* knockouts, which die hours after birth, m+62 ortholog deletion mice were born healthy at expected Mendelian ratios (Supplemental Figure 3.10a). The m+62 ortholog deletion mice also showed normal frequencies of B-cell progenitors in the fetal liver and mature B-lymphocytes in the adult peripheral blood (Supplemental Figure 3.10b, c). Other hematopoietic lineages were also observed at wild-type frequencies (Supplemental Figure 3.10c). BCL11A expression was unperturbed in the brain or sorted B cell precursors from E16.5 embryos (Supplemental Figure 3.10d). In contrast, there was substantial reduction in BCL11A levels in sorted E16.5 erythroid precursors (26% residual, $P<0.05$; Supplemental Figure 3.10d).

The m+62 ortholog deletion mice were bred to mice transgenic for the human β-globin cluster (β-YAC) to model the role of BCL11A in hemoglobin switching[131]. Unlike its fetal-stage expression in humans, in the mouse fetal liver transgenic human γ-globin is subject to intense repression (like an embryonic globin). *Bcl11a* is required for this early murine silencing of transgenic γ-globin at E14.5, although even in the absence of *Bcl11a*, γ-globin is ultimately repressed[135,136]. Fetal livers were evaluated between days E12.5 and E18.5 to monitor hemoglobin switching. Repression of human γ-globin and activation of human β-globin was markedly delayed in the m+62 ortholog deleted mice (Figure 3.5c). Heterozygous mice showed an intermediate γ-globin derepression phenotype, underscoring the dose-dependent inverse relationship between BCL11A and HbF level. These results indicate that targeting the erythroid enhancer of BCL11A *in vivo* results in erythroid-specific disruption of BCL11A expression and relaxed repression of γ-globin, unaccompanied by the obvious neurologic and immunologic toxicities seen in the BCL11A conventional knockout context.

DISCUSSION

We employed a novel application of CRISPR-Cas9 genome editing, saturating mutagenesis of noncoding elements *in situ*, to provide important insight into the organization and function of the BCL11A erythroid enhancer. Traditional tests of enhancer function rely on ectopic heterologous reporter assays and/or correlative biochemical features. Genome editing allows facile evaluation of the requirement of enhancer sequences within their endogenous chromatin context for appropriate gene regulation. As shown here, high-resolution high-throughput pooled tiling sgRNA screening reveals underlying enhancer sequence requirements approaching nucleotide resolution. A limitation to the resolution of this approach is the availability of NGG PAM sequences in a given region. We did not observe efficient editing by SpCas9 with NAG restricted sgRNAs (Supplemental Figure 3.2h , 3.6j). Recent studies have identified Cas9 orthologs and variants restricted by alternate PAM sequences, each capable of efficient genome editing[53,54,142]. This increased targeting range of Cas9 could allow increased resolution for *in situ* mutagenesis, particularly at sequences with paucity of NGG motifs. Alternatively, approaches reliant on homology-directed repair[143] could offer nucleotide resolution functional mutagenesis of noncoding sequences, though issues of efficiency, fidelity, and quantitative sensitivity would need to be considered. We suggest that our tiled pooled CRISPR screening approach could be readily adapted to the functional interrogation of numerous noncoding genomic elements.

In addition, these data demonstrate that apparent sequence conservation at the BCL11A enhancer masks underlying functional divergence. The mouse and human BCL11A erythroid composite enhancers share primary sequence homology, an erythroid enhancer chromatin signature, and syntenic intronic position relative to coding sequences. Moreover, both are required for erythroid expression of BCL11A and repression of embryonic/fetal globin genes. However, our high-resolution CRISPR mutagenesis analysis reveals divergence in the architecture of these enhancers. Of note, human BCL11A enforces the γ- to β-globin developmental switch around the time of birth. The timing and nature of these switches and the

globin genes themselves are distinct in primates as compared to nonprimate vertebrates that only exhibit a mid-gestation embryonic to adult switch[144]. Therefore it would seem plausible that critical regulatory mechanisms at *BCL11A* might differ between species (also see Supplementary Information).

The hemoglobin disorders represent one of the most common Mendelian inherited human conditions. The level of HbF is a key modifier of clinical severity of these diseases and BCL11A is the chief regulator of HbF level[144]. Naturally occurring genetic variation at the *BCL11A* enhancer is well-tolerated and associated with HbF level and β-hemoglobin disorder clinical severity. The work presented here offers a framework for therapeutic genome editing of the BCL11A enhancer for β-hemoglobin disorders. Enhancer disruption by individual sgRNAs in primary erythroid precursors results in substantial HbF induction. This approach may mitigate erythroid-specific growth disadvantages of complete BCL11A loss (Figure 3.2b). Furthermore erythroid enhancer disruption may spare BCL11A expression and function in nonerythroid contexts, such as B-lymphopoiesis (Supplemental Figure 3.10b-d). A challenge for the field is that it is not yet possible to accurately model HbF repression experimentally. However, individuals haploinsufficient for BCL11A due to microdeletions exhibit marked neurologic deficits, and elevated HbF beyond that seen in homozygotes for high-HbF common enhancer haplotypes[145,146]. Taken together, these data suggest that perturbation of critical sequences within the *BCL11A* enhancer defined here may result in HbF levels exceeding a clinical threshold required to ameliorate the β-hemoglobin disorders.

ACKNOWLEDGMENTS

This text has been reproduced from the following article published in the *Nature*:

Canver, MC., Smith, EC., Sher, F., Pinello, L., Sanjana, NE., Shalem, O., Chen, D., Schupp, P., Vinjamur, DS., Garcia, S., Luc, S., Kurita, R., Nakamura, Y., Fujiwara, Y., Maeda, T., Yuan, GC., Zhang, F., Orkin, SH., and Bauer DE. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature*. 527(7577): 192-7.

We thank J. Hughes and D. Higgs for assistance with analysis of ChIP-seq, R. Mathieu and the Boston Children's Hospital Hematology/Oncology-HSCI Flow Cytometry Research Facility for cell sorting, Z. Herbert and F. Abderazzaq at the Dana-Farber Cancer Institute Molecular Biology Core Facility and Center for Cancer Computational Biology respectively for sequencing, J. Doench for providing TALENs, C. Peng for advice with MEL reporter cell generation, F. Godinho and M. Nguyen for technical help with ESCs and transgenic mice, A. Dass, C. Lin, and S. Kamran for general technical assistance, C. Brendel and D. Williams for input regarding lentiviral transduction of HSPCs, J. Desimini for graphical assistance and J. Xu and G. Lettre for insightful discussions. M.C.C. is supported by F30DK103359-01A1. E.C.S. is supported by a Jane Coffin Childs Memorial Fund for Medical Research Fellowship. L.P. is supported by NHGRI Career Development Award K99HG008399. N.E.S. is supported by a Simons Center for the Social Brain Postdoctoral Fellowship and NIH NHGRI award K99-HG008171. O.S. is supported by a fellowship from the Klarman Family Foundation. S.L. is supported by a Leukemia & Lymphoma Society Fellow Award. T.M. is supported by NIH R01 A1084905. G.C.Y. is supported by NIH R01HL119099 and R01HG005085. F.Z. is supported by the NIMH (5DP1-MH100706) and NIDDK (5R01-DK097768), a Waterman award from the National Science Foundation, the Keck, McKnight, Damon Runyon, Searle Scholars, Merkin, Vallee, and Simons Foundations, and Bob Metcalfe. S.H.O. is supported by P01HL032262 and P30DK049216 (Center of Excellence in Molecular Hematology). D.E.B. is supported by an NIDDK Career

Development Award K08DK093705, Doris Duke Charitable Foundation Innovations in Clinical Research Award (2013137), and Charles H. Hood Foundation Child Health Research Award.
Computational tools and instructions for designing CRISPR-Cas9 sgRNA libraries for conducting non-coding screening can be found at the Zhang laboratory website http://www.genome-engineering.org.

D.E.B. conceived this study. N.E.S., O.S., and F.Z. conceived the pooled non-coding screening strategy using CRISPR-Cas9. M.C.C., N.E.S., O.S., F.Z., S.H.O., and D.E.B. designed and executed the pooled CRISPR screening strategy. E.C.S., F.S., Y.F., S.L., S.H.O., and D.E.B. designed, produced, and analyzed the transgenic mice. R.K. and Y.N. provided the HUDEP-2 cell line. M.C.C., F.S., T.M., S.H.O., and D.E.B. adapted the HUDEP-2 cell line as a model of globin gene regulation. M.C.C., F.S., D.C., P.S., D.S.V., and D.E.B. performed all experiments in cell lines. M.C.C., L.P., N.E.S., S.P.G., G.C.Y., F.Z., S.H.O., and D.E.B. analyzed the data. L.P., S.P.G. and G.C.Y. developed the HMM. M.C.C., S.H.O., and D.E.B. wrote the manuscript with input from all authors.

Chapter 4

Variant-aware saturating mutagenesis using multiple nucleases identifies regulatory elements underlying trait-associations of the HBS1L-MYB intergenic region

ABSTRACT

High-throughput, saturating in situ mutagenesis permits fine-mapping of function across DNA regions. Disease- and trait-associated variants from genome-wide association studies largely cluster in regulatory DNA. Here we demonstrate the use of multiple designer nucleases and variant-aware library design to interrogate trait-associated regulatory DNA at high resolution. *DNA Striker* offers a computational tool for the creation of saturating mutagenesis libraries with single or combinatorial nucleases with incorporation of variants from haplotype structure, whole-genome sequencing, or a custom list of variants. We applied this methodology to the HBS1L-MYB intergenic region, a locus associated with fetal hemoglobin levels and red blood cell traits. This approach identified four regulatory elements, including two that were previously identified and two novel elements. These data establish a high-throughput and high-resolution methodology to identify minimal functional sequences within large regions of disease- and trait-associated DNA.

INTRODUCTION

Genome-wide association studies (GWAS) are a powerful approach for identification of disease- and trait-associated variants. Greater than 90% of GWAS variants lie within regulatory DNA[97]. However, linkage disequilibrium often obscures the causal variant. Reliable methods to identify the causal variant and underlying functional sequence remain elusive. The clustered regularly interspaced short palindromic repeats (CRISPR)-based genome editing systems have emerged as highly efficient tools to study regulatory DNA. Targeted deletion provides a valuable tool to illuminate the function of regulatory DNA function through loss of function[67,147]. However,

targeted deletion is low throughput and has limited resolution[60]. Alternatively, the homology-directed repair (HDR) pathway can be exploited following cleavage by a designer nuclease to insert putative causal variants into endogenous DNA sequence using a customized extrachromosomal template. However, HDR to insert variants is also low-throughput and limited by its low efficiency. Furthermore, trait-associated variants may underestimate the effect of the underlying genetic element[67,147].

Saturating a region with insertions/deletions (indels) constitutes a powerful strategy to identify minimal functional sequences within regulatory DNA[147]. Saturating mutagenesis relies on pooled screening with every available single guide RNA (sgRNA) in a region to take advantage of the typical indel spectrum following non-homologous end joining (NHEJ) repair of 1-10 bp[48,60,68,74,93,147]. The ability to saturate a region with indels is a function of protospacer adjacent motif (PAM) availability. Moreover, genomic variants that attenuate sgRNA activity can reduce resolution through false negatives. We hypothesized that combining multiple nucleases with unique PAM sequences and incorporating variants into sgRNA library design would offer a high-throughput and high-resolution tool to interrogate trait-associated regulatory DNA.

METHODS

HUDEP-2 Karyotype

HUDEP clone 2 (HUDEP-2) was obtained from Nakamura and colleagues[122]. Karyotype analysis was performed at the Cytogenetics Laboratory at Tufts Medical Center.

HUDEP-2 Cell Culture

HUDEP-2 cells were expanded in SFEM (Stem Cell Technologies) supplemented with 100 ng/mL stem cell factor (R&D), 3 UI/mL erythropoietin (Amgen), 10^{-6} M dexamethasone (Sigma), 1 µ/mL of doxycycline (Sigma), and 2% penicillin-streptomycin (Thermo Fisher). HUDEP-2 cells

were differentiated in Iscove's Modified Dulbecco's Medium (IMDM) supplemented with 330 µg/mL holo-human transferrin (Sigma), 10 µg/mL recombinant human insulin (Sigma), 2 IU/mL heparin (Sigma), 5% human solvent detergent pooled plasma AB (Rhode Island Blood Center), 3 IU/mL erythropoietin (Amgen), 100 ng/mL human stem cell factor (SCF) (R&D), 1 µg/mL doxycycline (Sigma), 1% L-glutamine (Life Technologies), and 2% penicillin/streptomycin (Life Technologies).

HUDEP-2 NGG Cas9 and HUDEP-2 NGA Cas9 Cells

NGG Cas9 lentivirus was prepared as described below using LentiCas9-Blast (Addgene plasmid ID 52962). Cells were transduced with LentiCas9-Blast lentivirus and maintained with 10 µg/mL blasticidin (Sigma). The LentiCas9-Blast (Addgene plasmid ID 52962) plasmid was modified to include the VQR mutations as described in Kleinstiver et al[148]. NGA Cas9 lentivirus was prepared as described below using VQR-modified LentiCas9-Blast lentivirus. Cells were transduced with VQR-modified LentiCas9-Blast and maintained with 10 µg/mL blasticidin (Sigma).

NGG and NGA Cas9 Activity Reporters

To assess Cas9 activity, lentiviral reporters were used that included green florescent protein (GFP) and either an NGG-restricted or NGA-restricted sgRNA targeting GFP sequence. The NGG Cas9 activity reporter has been previously described[133]. In order to construct an NGA Cas9 activity reporter, pLentiGuide-Puro (Addgene plasmid ID 52963) was modified to express GFP and the NGA-restricted sgRNA sequence GTTCGAGGGCGACACCCTGG targeting GFP sequence. After transduction with the lentiviral reporters, successful transductants were selected with 1 µg/mL puromycin and incubated for 14 days to allow editing to occur.

Lentivirus Production

HEK293T cells were cultured with Dulbecco's Modified Eagle's Medium (DMEM) (Life Technologies) supplemented with 10% fetal bovine serum (FBS) (Omega Scientific) and 2% penicillin-streptomycin (Life Technologies). HEK29T were transfected at 80% confluence in 15 cm tissue culture treated petri dishes with 16.25 µg psPAX2, 8.75 µg VSV-G, and 25 µg of the lentiviral construct plasmid of interest using 150 µg of branched polyethylenimine (Sigma). Medium was refreshed 16-24 hours after transfection. Lentiviral supernatant was collected at 48 and 72 hours post-transfection. Viral supernatant was concentrated by ultracentrifugation (24,000 rpm for 2 hours at 4^{o}C; Beckman Coulter SW 32 Ti rotor).

Non-targeting sgRNA design

In order to design sgRNAs that do not target the human (hg19) and mouse (mm9) genomes, we first extracted all possible 20 bp sequences immediately preceding NG PAM motifs in both genomes. We created 5,000 random 20 bases sgRNA sequences that we compared to all 20 bp reference sequences. We calculated a targeting score dependent on the number and position of mismatches between both sequences using the methodology of Hsu et al[93]. The score ranges from 0 (non-targeting) to 1 (perfect match). We assigned a score of 0 to sequences with more than 4 mismatches. Reference sequences with score > 0 were considered potential off-targets. For each random guide, we derived an aggregated score from all possible off-targets, as per Hsu et al[93]:

$$S_{guide} = \frac{100}{100 + \sum_{i=0}^{n} S_{hit}(h_i)}$$

Where *n* is the number of potential off-target "hits", and $S_{hit}(h_i)$ is the targeting score of the possible off-target sequence h_i. In this situation, an aggregated score of 100 corresponds to no possible targets in the genome. Multiple off-targets or the presence of h_i-scoring off targets will

lower the score towards 0. We defined guides with an aggregated score > 90 as non-targeting (n=128).

Pooled CRISPR/Cas9 library design for high resolution, variant-informed functional mapping of the HBS1L-MYB intergenic region

The summit of every DNase hypersensitive site (DHS) within the HBS1L-MYB region (n = 98) was identified from fetal- and adult-derived $CD34^+$ subject to erythroid differentiation[67]. The regions of DHS summit +/- 200 bp were chosen for saturating mutagenesis based on previous work that suggested functional sequence was located at the peak of DNase hypersensitivity[147]. Using the *DNA Striker* tool, every 20-mer sequence upstream of an NGG or NGA PAM sequence on the sense or anti-sense strand was identified for each HBS1L-MYB region DHS as well as *BCL11A* exon 2, the core of the +58 DHS within the *BCL11A* enhancer[147], *HBS1L* exon 4, *MYB* exon 5 (Figure 4.1d-e). Phased variants within these region were taken from the 1,000 Genomes Project database in VCF file format (http://www.1000genomes.org/) using all individuals available by August 2015 (2,504 individuals; 5,008 alleles). Using the 1,000 Genomes variants, the variants feature within *DNA Striker* was used to identify sgRNAs altered by variants or new sgRNA resulting from PAM sequences created by variants. Variant-associated sgRNA were included in the library if variants were present at a frequency of ≥ 1% (Supplemental Fig. 4.10a,b). After filtering non-unique sgRNA, the NGG library was composed of 2,166 sgRNAs targeting HBS1L-MYB DHS, 176 variant-associated sgRNAs, 13 sgRNAs targeting HBS1L exon 4, 28 sgRNAs targeting MYB exon 5, 21 sgRNAs targeting the BCL11A enhancer +58 DHS core, 53 sgRNAs targeting BCL11A exon 2, and 128 non-targeting sgRNAs for a total of 2,585 sgRNAs. After filtering non-unique sgRNA, the NGA library was composed of 2,524 sgRNAs targeting HBS1L-MYB DHS, 186 variant-associated sgRNAs, 32 sgRNAs targeting HBS1L exon 4, 28 sgRNAs targeting MYB exon 5, 12 sgRNAs targeting the BCL11A enhancer +58 DHS core, 47 sgRNAs targeting BCL11A exon 2, and 128 non-targeting sgRNAs

for a total of 3,018 sgRNAs. Each of these 20-mer oligos were synthesized as previously described[123,124,147,149] and cloned using a Gibson Assembly master mix (New England Biolabs) into pLentiGuide-Puro (Addgene plasmid ID 52963). Plasmid libraries were deep sequenced to confirm representation (Supplemental Figure 4.10c).

Pooled CRISPR/Cas9 screen for high resolution, variant-informed functional mapping of the HBS1L-MYB intergenic region

HUDEP-2 cells with stable NGG or NGA Cas9 expression were transduced at low multiplicity with the corresponding NGG or NGA sgRNA library lentivirus pool while in expansion medium (NGG and NGA screens were performed independently). 10 µg/mL blasticidin (Sigma) and 1 µg/mL puromycin (Sigma) were added 24 hours after transduction to select for lentiviral library integrants in cells with Cas9. The screens for fetal hemoglobin expression in HUDEP-2 cells were performed as previously described[147]. Briefly, HUDEP-2 cells were differentiated and intracelllularly stained for HbF (clone HbF-1 with APC conjugation; Life Technologies). 0.2 µg HbF antibody was used per 500,000-5 million cells. An HbF-stained non-targeting sgRNA sample was used as a negative control to set a sorting gate for the HbF-high population (approximately top 5% of HbF-expressing cells). A corresponding percentage of cells from the HbF-low population were also sorted. After sorting the HbF-high and HbF-low pools, library preparation and deep sequencing was performed as previously described[84,147]. 6.6 µg of DNA per sample were submitted for Illumina MiSeq paired end sequencing with Nextera sequencing primers. Guide sequences present in the HbF-high and HbF-low pools were enumerated. HbF enrichment was determined by the log2 transformation of the median number of occurrences of a particular sgRNA in the HbF-high pool divided by the median number of occurrences of the same sgRNA in the HbF-low pool across the 3 biological screen replicates for each PAM-restricted library. Dropout scores were calculated by the ratio of normalized reads in the cells at end of experiment (average of reads in the HbF-high and HbF-low pools) to reads in the plasmid

pool for the median of the 3 biological screen replicates for each PAM-restricted library followed by log2 transformation. sgRNA sequences were mapped to the human genome (hg19). The plasmid library was deep sequenced to confirm representation using the same methodology. A quantile-quantile (Q-Q) plot was made with a line fitted through the first and third quantiles using MATLAB software.

Determination of PAM Distributions

Repeat-masked regions of the human genome (hg19) were removed. Non-repeat-masked repeats were parsed out separately to avoid creating false genomic junctions. PAMs were identified and the associated double strand break site for each potential sgRNA was determined. sgRNA with double strand break positions outside of these regions were excluded from analysis. Double strand break positions were compiled from sgRNA on both the plus and minus strands. The difference between adjacent genomic double strand break sites was calculated. Promoters (transcriptional start site +/- 2 kilobases), exons, and introns were determined from RefSeq annotations. Enhancer and DNase hypersensitive regions for GM12878, H1 hESC, HepG2, HMEC, HSMM, HUVEC, K562, NHEK, and NHLF cell lines were taken from publically available databases[150]. Repressed regions were used from previously published data[151].

Super Enhancer Analysis

Human H3K27ac ChIP-seq was obtained from a previously published dataset[104]. The ROSE algorithm was used to perform super enhancer analysis[107].

Variant Datasets

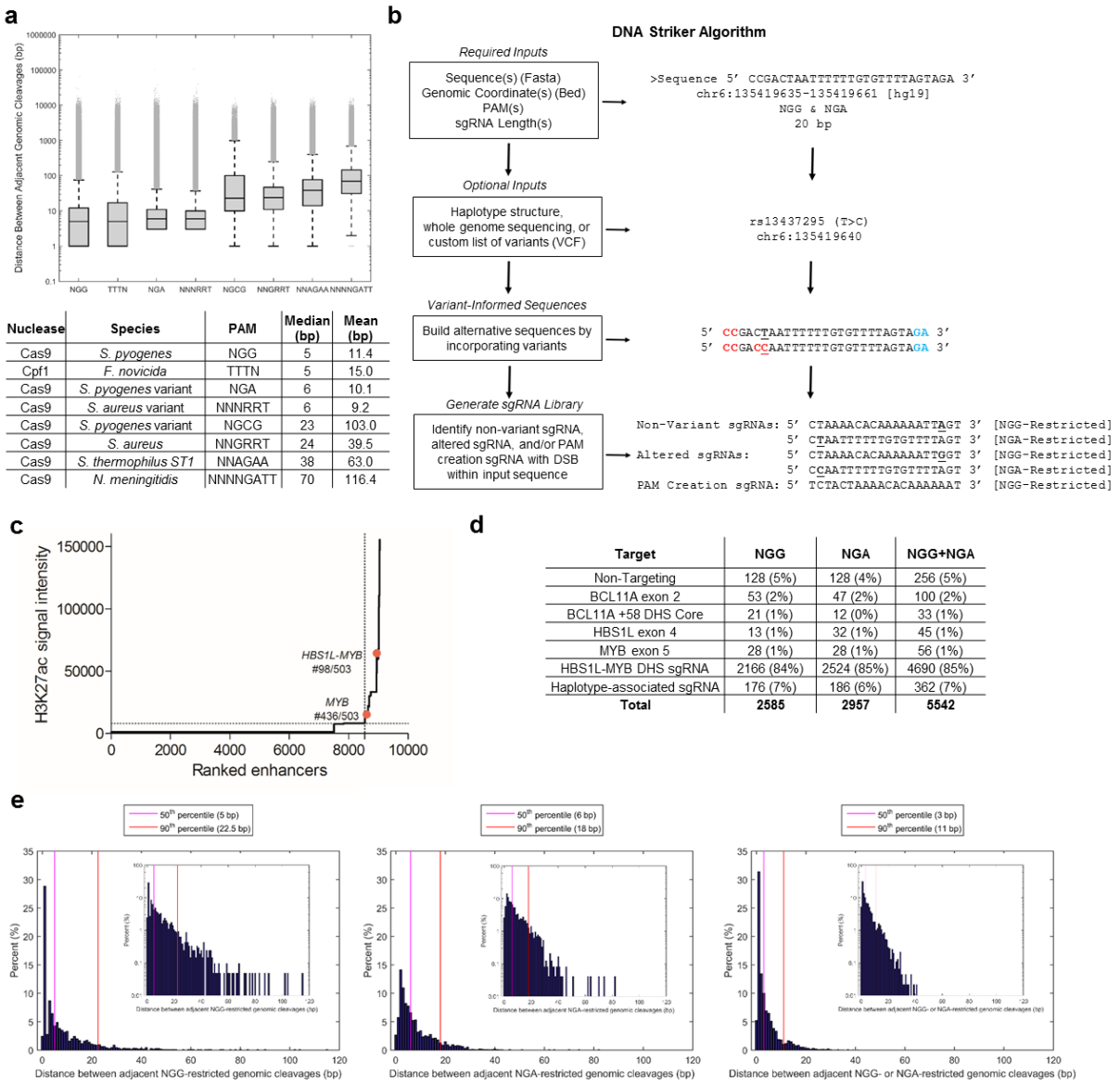
Representative *DNA Striker* output plots were generated for chr6:135282411-135282852 (hg19) using haplotype data generated as described above, publically available whole genome

sequencing in VCF format from NA12878, and a custom list of variants of SNPs. The custom list of SNPs included variants with a minor allele frequency of ≥1% in this region. A VCF of this custom list of variants was obtained from the dbSNP database.

RESULTS

Distribution of PAM sequences in the genome and outline of the *DNA Striker* algorithm

A variety of CRISPR-associated nucleases with unique PAM recognition sequences have been used for genome editing[48,50,53,54,74,148,152,153]. The degree of saturation of each nuclease is dependent on minimizing genomic distance between potential adjacent cleavages. The levels of saturation for each PAM in the genome vary (Figure 4.1a; Supplemental Figures 4.1-4.2). Given the sequence-dependence of PAM availability, regional-based variation in degree of saturation for each nuclease is observed in DNase I hypersensitivity sites (DHS), enhancers, and repressed regions as well as genes (promoters, exons, introns) (Supplemental Figures 4.3-4.7). We hypothesized that the usage of multiple nucleases in combination with variant-aware saturating mutagenesis library design can optimize resolution and reliability to identify trait-associated regulatory elements (Figure 4.1b). We created *DNA Striker* as a MATLAB-based computational tool to design saturating mutagenesis libraries using single or combinatorial designer nucleases as well as provide alternative sgRNA based on haplotype structure, whole-genome sequencing, or a custom list of variants. The algorithm for *DNA Striker* is summarized in Figure 4.1b. Briefly, uploaded DNA sequence(s) (fasta format) are analyzed for all PAM(s) sequences requested by the user using a sliding window approach. Users must provide genomic coordinates for each sequence (bed format) and must choose the sgRNA length for each PAM sequence in the library given that optimal sgRNA vary for different CRISPR-associated nucleases[48,50,53,54,74]. Variant-aware sgRNA library design involves identifying sgRNA altered by variants and novel sgRNA resulting from PAM sequences created by the presence of variants (Figure 4.1b).



Nuclease	Species	PAM	Median (bp)	Mean (bp)
Cas9	*S. pyogenes*	NGG	5	11.4
Cpf1	*F. novicida*	TTTN	5	15.0
Cas9	*S. pyogenes* variant	NGA	6	10.1
Cas9	*S. aureus* variant	NNNRRT	6	9.2
Cas9	*S. pyogenes* variant	NGCG	23	103.0
Cas9	*S. aureus*	NNGRRT	24	39.5
Cas9	*S. thermophilus ST1*	NNAGAA	38	63.0
Cas9	*N. meningitidis*	NNNNGATT	70	116.4

Target	NGG	NGA	NGG+NGA
Non-Targeting	128 (5%)	128 (4%)	256 (5%)
BCL11A exon 2	53 (2%)	47 (2%)	100 (2%)
BCL11A +58 DHS Core	21 (1%)	12 (0%)	33 (1%)
HBS1L exon 4	13 (1%)	32 (1%)	45 (1%)
MYB exon 5	28 (1%)	28 (1%)	56 (1%)
HBS1L-MYB DHS sgRNA	2166 (84%)	2524 (85%)	4690 (85%)
Haplotype-associated sgRNA	176 (7%)	186 (6%)	362 (7%)
Total	**2585**	**2957**	**5542**

Figure 4.1: Degree of saturation is a function of the availability of PAM sequences. a, Distances between adjacent genomic cleavages to assess genome-wide PAM availability and distribution. For each box-and-whisker plot, the three lines of box represent the 25th, 50th and 75th percentile. The upper and lower whiskers represent the 99th and 1st percentile, respectively. Outliers, defined as above the 99th percentile or below the 1st percentile are plotted as individual points. Lower whiskers are omitted if the 1st percentile is 0. **b**, Description of the DNA Striker algorithm. **c**, Ranked enhancers by H3K27ac signal intensity from primary human adult erythroid precursors. Super-enhancers occur in the region above the horizontal dotted line and to the right of the vertical dotted line. Two super enhancers within the HBS1L-MYB intergenic region are indicated. **d**, Library composition for NGG-restricted sgRNA-only, NGA-restricted sgRNA-only, as well as NGG- and NGA-restricted sgRNA together. **e**, For the HBS1L-MYB intergenic region DHS, the degree of saturation of using NGG-only (*left panel*), NGA-only (*middle panel*), and NGG and NGA combined (*right panel*).

Variant analysis for whole genome sequencing or custom list of variants (both in VCF format) occurs by creating multiple versions of the sliding window: the non-variant version,

versions with each variant in the window inserted in isolation, and all combinations of up to three variants in each window. Variant analysis for haplotype data (VCF format) occurs by creating each individual allele present in the haplotype data provided. Analysis completes by outputting the full library design and two output figures assessing the degree of saturation of the uploaded sequence(s) (Supplemental Figure 4.8).

Saturating mutagenesis library design

GWAS, quantitative trait loci (QTL), and other human genetic studies of fetal hemoglobin (HbF) level (or the related trait F-cell number) have implicated three loci, including the HBS1L-MYB interval[154–161]. In addition to HbF, MYB has also been associated with various erythroid traits[162–167]. These associations have been suggested to reflect changes in the expression of *MYB* due to distant variants localizing kilobases away and approximately equidistant to the *HBS1L* gene. The density of trait-associated single nucleotide polymorphisms (SNPs) reside in an 83 kilobase super enhancer (Figure 4.1c, 4.2a, b). Recent studies have utilized lineage-restricted expression patterns, clustering of erythroid transcription factor binding sites affecting MYB expression, and chromatin capture to suggest that HbF-associated variants modulate MYB expression by altering two regulatory elements -71 and -84 kb upstream of the *MYB* transcriptional start site (TSS)[159].

The HBS1L-MYB region is comprised of 98 DNase hypersensitive sites (DHSs) as identified from fetal- and adult-derived $CD34^+$ subject to erythroid differentiation[67]. In order to interrogate this intergenic region in a comprehensive fashion, the regions of each DHS summit (peak of DNase sensitivity) +/- 200 bp were chosen for saturating mutagenesis based on previous work that suggested functional sequence localizes to the peak of DNase hypersensitivity[147]. Using *DNA Striker*, we designed a high resolution saturating mutagenesis library consisting of all NGG- and NGA-PAM restricted sgRNA given their high degree of saturation in DHS (Figure 4.1d; Supplemental Figures 4.3, 4.7). The median and 90^{th} percentile

gap distance between adjacent genomic cleavages using NGG Cas9 was 5 bp and 22.5 bp, respectively while it was 6 bp and 18 bp for NGA Cas9 (Figure 4.1e). The combination of using both NGG and NGA Cas9 nucleases led to a reduction in the median and 90th percentile gap between adjacent genomic cleavages to 3 bp and 11bp, respectively. Furthermore, use of both nucleases reduced the maximum gap size from 115 bp for NGG and 82 bp for NGA Cas9 to a maximum of 41 bp for the combined setting. Therefore, the usage of two nucleases resulted in higher resolution by reducing the 50th and 90th percentile of distances between adjacent genomic cleavages as well as reducing the maximum gap between adjacent cleavages (Supplemental Figure 4.9).

The library was comprised of every 20-mer sequence upstream of an NGG or NGA PAM sequence on the sense or anti-sense strand within the HBS1L-MYB region DHS as well as *BCL11A* exon 2, the core of the +58 DHS within the *BCL11A* enhancer[147], *HBS1L* exon 4, and *MYB* exon 5 (Figure 4.1d, e). To construct a variant-informed library, phased variants within these regions were taken from the 1,000 Genomes Project database and incorporated into the sgRNA design by *DNA Striker* to identify potential altered sgRNA and novel sgRNA resulting from variant-induced PAM creation (Figure 4.1b). Variants present within the database with a frequency ≥1% were chosen for inclusion in each library (Figure 4.1b, d; Supplemental Figure 4.10a, b). Both NGG- and NGA-restricted sgRNA libraries were batch cloned into lentiviral constructs for screening (Supplemental Figure 4.10c).

Functional saturating mutagenesis screens using NGG and NGA Cas9s

To demonstrate specificity and efficiency of *S. pyogenes* NGG Cas9 and *S. pyogenes* variant NGA Cas9[148], we used Cas9 reporter constructs that delivered GFP as well as either an NGG-restricted or NGA-restricted sgRNA targeting GFP. Cells stably expressing NGG Cas9, NGA Cas9, or no Cas9 were transduced with the reporter construct at low multiplicity and selected for 14 days. The analysis demonstrated that the NGG and NGA Cas9 proteins were both specific

and efficient nucleases as NGG Cas9 only led to significant GFP reduction with an NGG-restricted sgRNA, and vice versa (Figure 4.2a). The HUDEP-2 erythroid cell line was used to

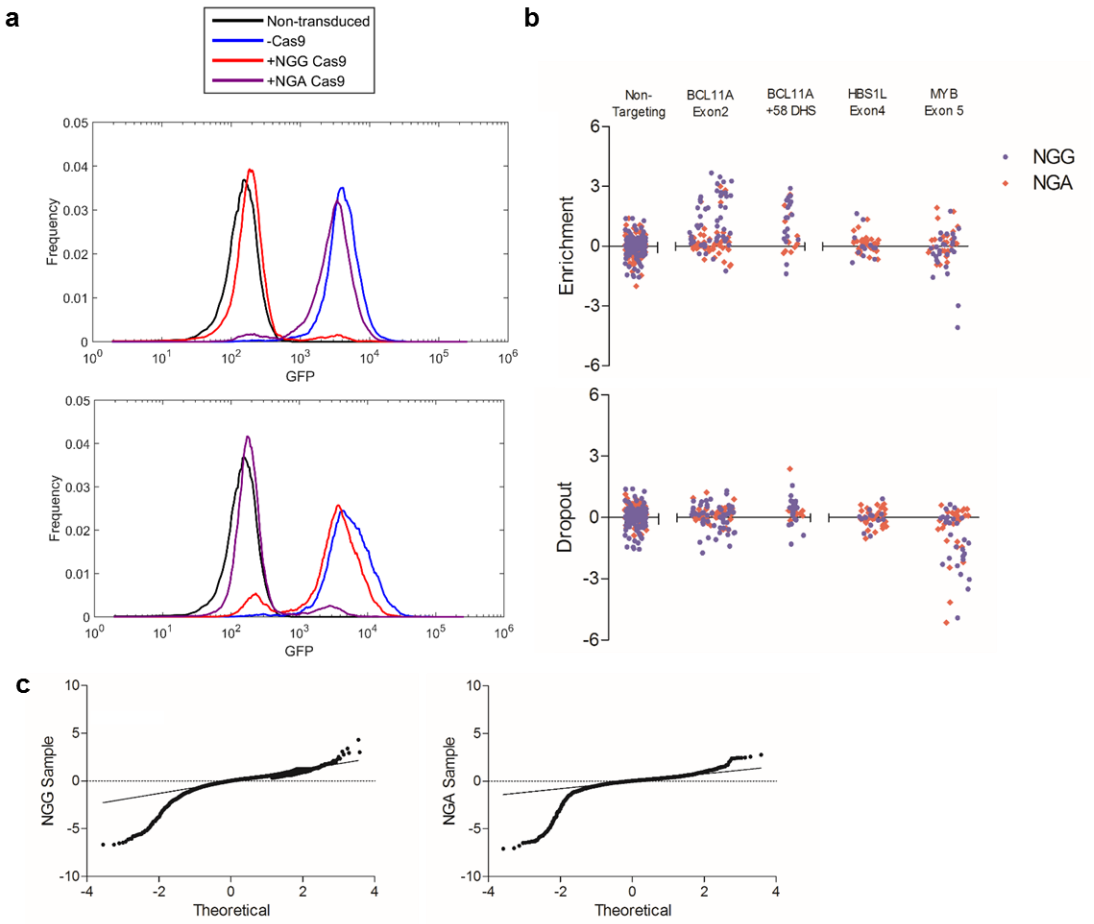


Figure 4.2: Pooled saturating mutagenesis screening of the HBS1L-MYB region using NGG- and NGA Cas9s. a, Cells stably expressing NGG (*red*), NGA (*purple*), or without Cas9 (*blue*) were transduced with a Cas9 activity reporter, which contained GFP and either an NGG- or NGA-restricted sgRNA. A non-transduced sample (*black*) was included as a negative control. **b**, Mapping enrichment and dropout scores to associated genomic loci. Non-targeting sgRNA are pseudo-mapped with 5-bp spacing. **c**, Quantile-quantile plots of NGG and NGA dropout scores.

examine HbF expression as previously described[122,147]. Briefly, HUDEP-2 cells stably expressing NGG or NGA Cas9 were transduced at low multiplicity with the associated NGG-restricted or NGA-restricted library. Cells were expanded, differentiated, sorted for high and low HbF-expression, and deep sequenced to enumerate sgRNA present within the HbF-high and HbF-low pools. Three biological replicates were performed for both libraries. Surprisingly, there was a lack of enrichment of both *HBS1L* exon 4- and *MYB* exon 5-targeted sgRNA (Figure 4.2b)

while both positive controls, *BCL11A* exon 2 targeted and +58 DHS within the *BCL11A* enhancer targeted sgRNA, enriched in the HbF-high pool. However, sgRNA targeting MYB showed a preponderance to dropout of the screen consistent with MYB's known essential role in erythropoiesis. *HBS1L* and *BCL11A* +58 DHS targeted sgRNA were not underrepresented, whereas *BCL11A* exon 2 sgRNA showed modest dropout consistent with previous findings[147] (Figure 4.2b). Upon dropout analysis of the NGG and NGA libraries, we determined that the majority of sgRNA in the library did not dropout, suggesting a neutral effect on erythropoiesis (Figure 4.2c). Notably, both libraries identified specific sgRNAs with significant dropout (Figure 4.2c). The presence of multiple colocalizing sgRNA in an *in situ* saturating mutagenesis screens has been previously shown to identify minimal functional sequences[147]. Upon mapping these sgRNAs to their associated genomic loci, the significant dropout sgRNAs colocalized to four discrete loci termed -126, -83, -71, and -7 as their distance in kilobases from the MYB TSS for both the NGG- and NGA-restricted libraries (Figure 4.3a-c). Of note, the -83 and -71 DHS fall within an annotated super enhancer region. These four identified sites strongly suggest regulatory potential and likely suggest control of *MYB* expression levels.

DISCUSSION

The functional sequences responsible for the association of the HBS1L-MYB intergenic region to HbF and other erythroid traits have been unknown due to a lack of methods to interrogate the function of trait-associated non-coding sequences in a high-throughput manner. We propose that high-resolution, variant-informed CRISPR-based saturating mutagenesis can provide a powerful tool with which to identify functional regions within variant-decorated regulatory DNA. Notably, previous studies had focused on two functional regions, -84 and -71[159]. Our approach identified these two known DHS within one kilobase and also implicated two additional loci (-126 and -7). Of interest, these data suggest that the -83 DHS contains functional sequences as

opposed to the -84 DHS, which highlights the enhanced resolution of this technique as opposed to methods such as chromosomal conformational capture and motif analysis. Moreover, the

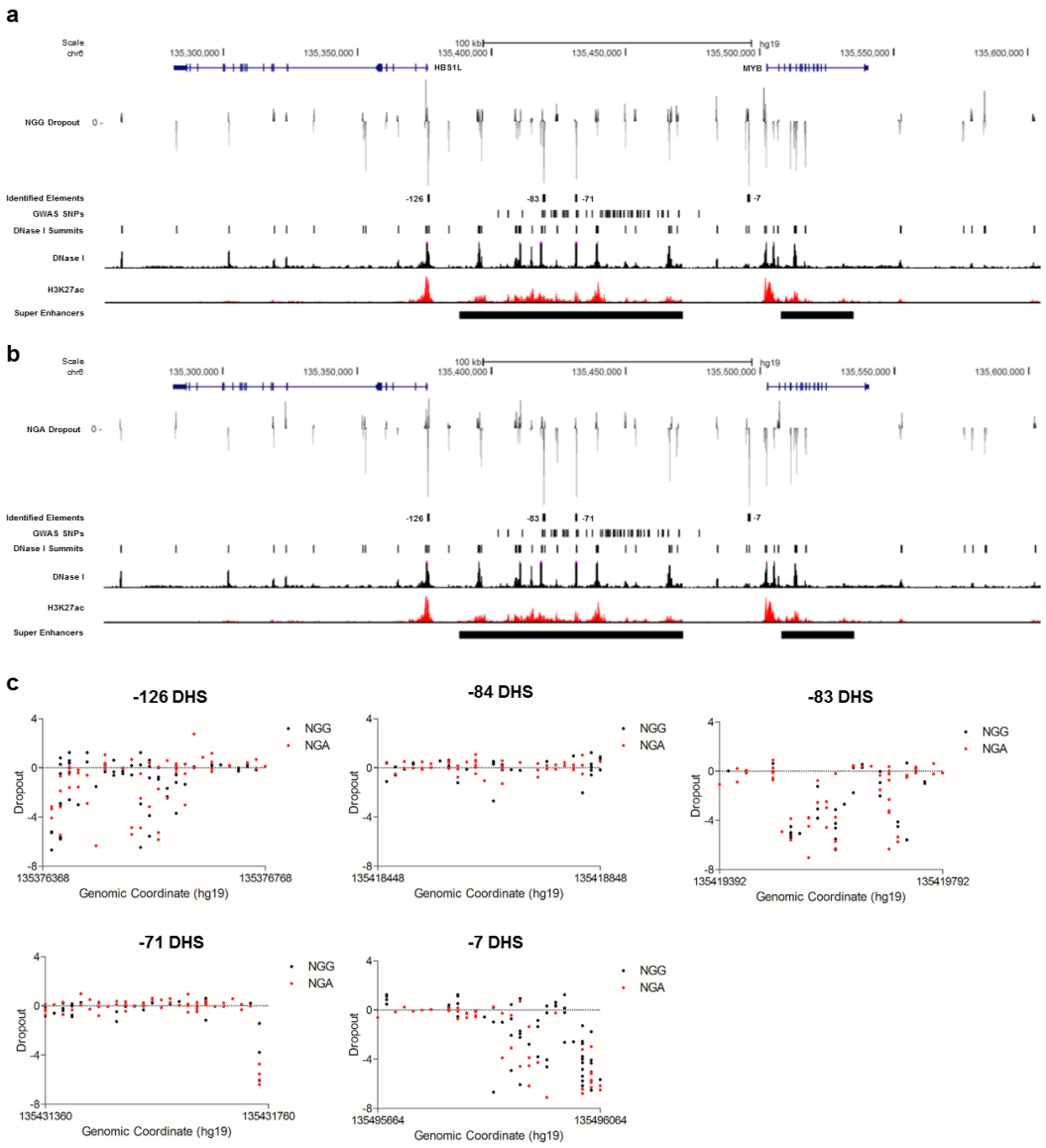


Figure 4.3: Mapped NGG- and NGA-restricted sgRNA to associated genomic loci identifies four functional elements. a, Mapping NGG dropout scores to associated genomic loci identifies four identified functional elements, -126, -83, -71, and -7 are indicated. **b**, Mapping NGA dropout scores to associated genomic loci identifies four identified functional elements, -126, -83, -71, and -7 are indicated. **c**, NGG- and NGA-restricted sgRNA for each of the identified loci. The -84 DHS has been previously associated with altering MYB expression[20].

HBS1L-MYB intergenic region on chromosome 6 is trisomic in HUDEP-2 cells (Supplemental Fig. 4.11). The ability to obtain interpretable data in a trisomic context underscores the robustness of this approach to interrogate sequence function even in the setting of aneuploidy.

These data establish high-resolution, variant-informed saturating mutagenesis as a powerful and high-throughput approach for identification of functional sequences in disease- and trait-associated regulatory DNA.

ACKNOWLEDGMENTS

This text has been reproduced from the following manuscript in preparation: **Canver, MC.**, Pinello, L., Lessard, S., Stern, E., Needleman, A., Chen, DD., Vinjamur, DS., Kurita, R., Nakamura, Y., Lettre, G., Yuan, GC., Bauer, DE. & Orkin, SH. (2016). Variant-aware saturating mutagenesis using multiple nucleases identifies regulatory elements underlying trait-associations of the HBS1L-MYB intergenic region.

We thank Z. Herbert, M. Berkeley, and M. Vangala at the Dana-Farber Cancer Institute Molecular Biology Core Facility for sequencing and members at the Hematologic Neoplasia Flow Cytometry and the Flow Cytometry Core facilities at the Dana-Farber Cancer Institute for cell sorting. M.C.C. is supported by a National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Award (F30DK103359-01A1). L.P. is supported by a National Human Genome Research Institute (NHGRI) Career Development Award (K99HG008399). S.L. is funded by a Canadian Institutes of Health research Banting doctoral scholarship. E.N.S. is supported by a Hematology Opportunities for the Next Generation of Research Scientists (HONORS) award from the American Society of Hematology. G.L. is funded by the Canada Research Program, the Montreal Heart Institute Foundation, and the Canadian Institute of Health Research (MOP136979 and MOP123382). G.C.Y. is supported by awards from the National Heart, Lung, and Blood Institute (NHLBI) (R01HL119099 and R01HG005085). D.E.B. is supported by an NIDDK Career Development Award (K08DK093705), Doris Duke Charitable Foundation Innovations in Clinical Research Award (2013137), and Charles H. Hood Foundation Child Health Research Award. S.H.O. is supported by an award from the NHLBI

award (P01HL032262) and an award from the NIDDK (P30DK049216, Center of Excellence in Molecular Hematology).

AUTHOR CONTRIBUTIONS

M.C.C., D.E.B., and S.H.O. conceived this study. M.C.C. developed the *DNA Striker* computational tool. M.C.C., E.S., A.N., D.D.C., D.S.V. performed the experiments. R.K. and Y.N. provided the HUDEP-2 cell line. M.C.C., L.P., S.L., G.L., G.C.Y. performed computational data and statistical analysis. D.E.B. and S.H.O. supervised this work. M.C.C., D.E.B., and S.H.O. wrote the manuscript.

Chapter 5

Conclusion

GENERAL STRATEGIES FOR SCD AND β-THALASSEMIA THERAPEUTIC GENOME EDITING

Correction of underlying genetic defects

The most appealing and theoretically straightforward application of genome editing for monogenic disorders is correction of a mutant DNA sequence and in that manner preserving all intrinsic regulatory mechanisms acting on the gene of interest. Precise gene correction relies on HDR from an extrachromosomal template containing a wild-type gene sequence. Typically, the frequency of HDR is relatively low, and particularly low in $CD34^+$ hematopoietic stem and progenitor cells (HSPCs) as discussed below. However, gene correction strategies may benefit from mixed chimerism allogeneic transplant studies suggesting that low levels of chimerism can produce clinical benefit[168]. Clinical development of such strategies requires optimizing efficiency and safety of correcting the sickle mutation, whereas the diverse spectrum of β-thalassemia point mutations and deletions necessitates optimization for each unique genetic target, a significant challenge for clinical translation.

Gene editing for reactivation of HbF in SCD and β-thalassemia

Elevated HbF is beneficial in SCD and β-thalassemia. Targets for manipulation include sequences lying within the β-globin cluster or within the genes encoding transcriptional regulators of globin gene expression (Figure 5.1 and Table 5.1). Depending on the target, editing would rely on HDR or NHEJ. Suitability of each target relates to the ease with which the desired gene modifications can be generated and the extent to which the modifications reactivate HbF expression. In SCD the goal is to induce sufficient HbF to prevent HbS

polymerization. In β-thalassemia, the aim is to replace deficient β-globin and thereby reduce globin chain imbalance.

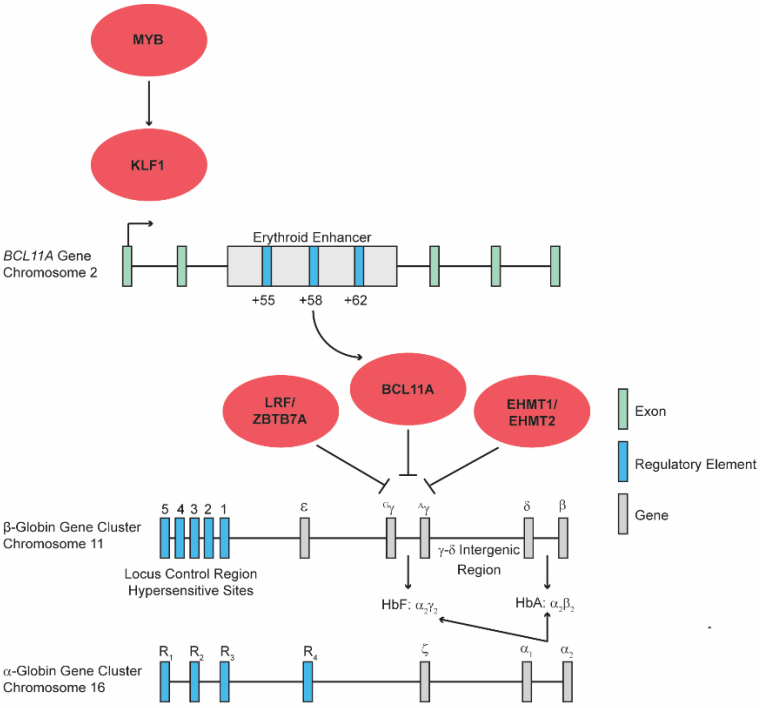


Figure 5.1: Network of potential targets for genome-editing based therapy of the β-globin disorders. Therapeutic genome editing strategies rely on genetic correction through the HDR pathway or genetic disruption through the NHEJ pathway. Genetic correction/repair strategies involve direct modification of the β-globin gene cluster through (1) correction of the sickle mutation in the β-globin gene or (2) insertion of the HPFH-associated SNPs into the ^{G}γ or ^{A}γ promoters. Genetic disruption strategies involve targeted disruption of (1) *BCL11A* coding sequence, (2) the minimal critical sequences in the +58 DHS within the erythroid-specific *BCL11A* enhancer, (3) the HbF-associated sequences within the ^{A}γ-δ intergenic region, or (4) other genes with a known role in γ-globin regulation such as MYB, KLF1, LRF/ZBTB7A, or EHMT1/EHMT2.

POSSIBLE TARGETS FOR SCD AND β-THALASSEMIA THERAPEUTIC GENOME EDITING

Genetic correction of the SCD and β-thalassemia mutations

Classical gene targeting approaches have been used to repair the SCD mutation in embryonic stem cells[169], but this approach cannot be applied to CD34^{+} HSPCs due to low efficiency and the necessity to isolate and propagate faithful recombinants. Correction of genetic defects in cultured cells with an engineered nuclease and a donor repair template has been achieved for multiple disorders, including cystic fibrosis[86,87,170,171], Duchenne's muscular dystrophy[172,173], and other diseases[71,174–177]. Gene correction for SCD and β-thalassemia has also been

accomplished in a laboratory setting[169,178–181]. Of note, a recent study reported correction of an SCD allele at nearly 20% gene in $CD34^+$ HSPCs upon delivery of a repair template via

Table 5.1: Potential targets for therapeutic genome editing for the β-globin disorders.

Target	Repair Strategy	Efficiency	Advantages/Disadvantages
Repair of SCD allele	HDR	Low-moderate	Single target allele; inadvertent generation of β-thalassemia alleles
Repair of β-thalassemia allele	HDR	Low-moderate	Heterogeneous target alleles
Recreation of non-deletional HPFH	HDR	Low-moderate	Inadvertent generation of γ-null alleles; identified HPFH patients support mutation tolerance/clinical benefit
Recreation of deletional HPFH	NHEJ	Low-moderate	Insufficient efficiency of targeted deletion; identified HPFH patients support mutation tolerance/clinical benefit
Other targets in β-globin cluster	NHEJ	-	Targets unknown
BCL11A	NHEJ	High	HSC/B-cell dysfunction due to BCL11A requirement; haploinsufficient patients have significant HbF induction
BCL11A enhancer	NHEJ	High	Erythroid-specific BCL11A loss; haploinsufficient patients have significant HbF induction
α-globin	NHEJ	Moderate-high	Balance α-β chains; inadvertent generation of α-thalassemia cells
KLF1	NHEJ	High	Broad role in cell proliferation and cellular development
MYB	NHEJ	High	Broad role in cell proliferation and cellular development
LRF/ZBTB7A	NHEJ	High	Broad role in cell proliferation and cellular development
EHMT1/EHMT2	NHEJ	High	Role in hematopoiesis unknown
LIN28B pathway	NHEJ	High	Role in hematopoiesis unknown

integration-deficient lentivirus or by DNA oligonucleotide electroporation in the presence of a β-globin targeted ZFN. Similar levels of correction were observed in bone marrow cells isolated from SCD patients[178]. Despite successful HDR in bulk cells *in vitro*, the levels of HDR were reduced in the spleen and bone marrow of transplanted immunodeficient mice, suggesting that HDR within long-term engrafting hematopoietic stem cells (HSCs) was far less efficient than in

downstream progenitors. Another study reported improved rates of HDR in repopulating cells via electroporation of ZFN mRNA in conjunction with an adeno-associated virus (AAV) donor repair template[182]. Studies using AAV in conjunction with megaTALs[183], transcription activator-like effectors coupled to a sequence specific homing endonuclease, demonstrated ~14% rates of HDR in $CD34^+$ HSPCs[184]. While megaTALs may enhance HDR through generation of 3' DNA overhangs in HSPCs, the rate of HDR in repopulating HSCs has not been examined.

The relative efficiency of HDR versus NHEJ is critical to potential use of gene editing for gene correction. High rates of NHEJ-mediated indel formation are suboptimal for clinical translation of β-globin gene correction as the process creates the possibility of disruption of β-globin production and inadvertent generation of β-thalassemia alleles. Another consideration is that mutagenesis has also been observed in the highly homologous δ-globin gene in β-globin gene correction experiments, which may result in deletions and rearrangements affecting β-globin that may be difficult to detect by standard PCR-based genotyping approaches[178].

It is possible that small molecules that enhance HDR and/or inhibit NHEJ may improve the efficiency of gene correction within $CD34^+$ HSPCs, so long as they do not impair cell engraftment capability[185,186]. NHEJ is the dominant pathway in G1, S, and G2 phases of the cell cycle, whereas HDR preferentially occurs during late S-phase and G2 phase when sister chromatid templates become available[187]. Since HSCs, the rare long-term repopulating cells within $CD34^+$ HSPC preparations, are largely quiescent, HDR is not favored. These observations are supported by the roles of BRCA1, PALB2, and BRCA2 in DSB repair. BRCA1 creates single strand DNA through end resection and interacts with PALB2 to recruit BRCA2 and RAD51 to mediate HDR at sites of DSB. Identification of the cell cycle's role in suppressing BRCA1 in the G1 phase supports the dominance of NHEJ repair in quiescent cells[188]. However, restoration of the BRCA1-PALB2 interaction during the G1 phase can support HDR. Therefore, it may be possible to enhance HDR in quiescent HSCs through modulation of the BRCA1-PALB2-BRCA2 pathway[188]. Moreover, one study demonstrated enhanced rates of HDR in

HEK293T and non-hematopoietic primary cells through cell cycle synchronization to achieve nuclease-mediated cleavage during the optimal portions of the cell cycle for HDR[187]. However, triggering proliferation in HSCs tends to impair their ultimate repopulating potential. Whether expansion of HSPC populations with small molecules such as SR1[189,190] or UM171[191] will allow for improved HDR efficiencies with concomitant retention of stem cell activity *in vivo* is as yet unknown.

Modification of the β-globin locus to recreate hereditary persistence of fetal hemoglobin

As would be anticipated from the existence of rare hereditary persistence of fetal hemoglobin (HPFH) alleles, genome wide association studies (GWAS) have linked the β-globin cluster itself to HbF levels[154,155,192–195]. This corroborated previous human genetic studies that identified HPFH patients with elevated HbF levels resulting from large deletions within the β-globin cluster[144,196,197]. Re-creating the larger deletional HPFH alleles is impractical given their large size[60]. However, opportunities may exist for targeting discrete regions of the β-globin gene cluster by NHEJ. Comparison of large deletions in the cluster that generate either HPFH or δβ-thalassemia phenotypes has implicated sequences in the ^{A}γ-δ intergenic region as harboring silencers of γ-gene expression. Notably, study of three families with overlapping deletions in the β-globin cluster identified a 3.5 kilobase region between the ^{A}γ and δ genes that may be essential for γ-globin repression. Additional, indirect support was derived from chromatin immunoprecipitation-PCR experiments that suggest BCL11A binding within this region[134,196,198–200]. At present, the optimal sequences in the cluster amenable for targeted deletion by editing and NHEJ have not been identified.

Several point mutations or small deletions in the ^{A}γ or ^{G}γ-globin gene promoters lead to persistence of HbF into adult life. HbF levels in heterozygotes with these nondeletional HPFH mutations may be as high as 30%[199,201–203]. One of the strongest HPFH alleles (-175 T>C in the ^{A}γ promoter) was recently created in cultured K562 cells with TALENs. Increased γ-globin

production resulted, most likely through *de novo* generation of a TAL1 binding site that facilitated increased chromatin looping between the $^{A}\gamma$ promoter and the locus control region[204]. An HPFH allele with a small deletion in the $^{A}\gamma$ promoter was re-created in CD34^{+} HSPCs with a sgRNA and Cas9 expression, presumably due to microdeletion of a repeated sequence[205]. Therapeutic genome editing to generate HPFH mutations is an attractive strategy as the effects of these mutations are known through study of families with these rare beneficial alleles. The approach, however, faces many of the same challenges as precise gene correction, given the apparent dominance of the NHEJ pathway at the expense of HDR efficiency in HSCs.

BCL11A targeting

BCL11A gene disruption

The GWAS-implicated transcription factor BCL11A is a validated repressor of HbF[134,135,154,155,192–195]. Erythroid-lineage Bcl11a knockout in a mouse model of SCD led to pancellular HbF induction and phenotypic correction of a mouse model of sickle cell disease[136]. Haploinsufficient patients with microdeletions within the BCL11A locus have significant neurocognitive phenotypes as well as elevated HbF at levels near or above therapeutic thresholds[206,207]. In principle, the genetic knockout of BCL11A by targeting *BCL11A* coding sequence in order to create frameshift null alleles represents a potential therapeutic strategy. Roles of *BCL11A* in non-hematopoietic lineages including the neural lineage[140,208], pancreatic progenitors[209], and the breast epithelium[210] would not be problematic upon modification of *BCL11A* in autologous CD34^{+} HSPCs. However, this strategy is limited by extra-erythroid roles *of BCL11A* in the hematopoietic system, including its requirement for B-cell development[139,140,211,212] and HSC function[213,214]. These roadblocks might be circumvented by use of erythroid restricted expression of genome editing components. A variation of this approach involves erythroid-specific, shRNA-mediated knockdown of *BCL11A* expression,

which is under development as a gene therapy strategy[215]. Delivery of genome editing tools stably to $CD34^+$ HSPCs would be inadvisable due to potential insertional mutagenesis[216] as well as elevated risk of off target mutagenesis over time. Furthermore, the effects of long-term expression of ZFNs, TALENs, or CRISPR/Cas9 on $CD34^+$ HSPCs are unknown.

BCL11A gene enhancer

Recent fine mapping of HbF-associated GWAS variants led to the identification of a developmental stage-specific, erythroid-restricted 12-kb region bearing a characteristic enhancer chromatin signature. This enhancer region is composed of three DNaseI hypersensitive sites (DHS), termed +55, +58, and +62 as their distance in kilobases from the *BCL11A* transcriptional start site. Deletion of the orthologous element in a murine erythroid cell line resulted in a complete loss of BCL11A at both the RNA and protein levels while expression was spared in a B cell line with the same deletion[67]. Subsequent deletion studies demonstrated a similar requirement for this element for *BCL11A* expression in human erythroid cells[147,217].

BCL11A enhancer targeting has several distinct advantages over coding sequence disruption: (1) GWAS studies have demonstrated that variation in the *BCL11A* enhancer is associated with elevated HbF levels and is both common and well-tolerated[67]. (2) Targeted deletion of this element in a human cell line leads to loss of *BCL11A* expression and subsequent HbF induction nearly comparable to *BCL11A* null clones[147]. (3) Targeted deletion of the murine +62 DHS within the *Bcl11a* erythroid enhancer results in delayed hemoglobin switching sparing expression in the brain and non-erythroid hematopoietic lineages[147]. The +62 DHS knockout mice were viable and born in normal Mendelian ratios as compared to *Bcl11a*$^{-/-}$ knockout mice that are perinatal lethal likely to do neural defects[135,147]. These results further highlight the erythroid specificity of this element *in vivo*[147]. (4) Targeting the *BCL11A* enhancer has been shown to be better tolerated even within the erythroid lineage as compared to

targeting *BCL11A* coding sequence, suggesting a residual low level of BCL11A present after enhancer targeting is insufficient to repress γ-globin, but promotes cellular fitness[147,217].

Therefore, an alternative approach to targeting *BCL11A* coding sequence might be targeted deletion of the 12 kilobase *BCL11A* erythroid enhancer[67]. However, while targeted deletions from ~1 kb to 1 Mb have been demonstrated to occur at an appreciable frequency, these are unlikely to occur at a sufficient frequency at clinical scale with current genome editing technologies due to competing outcomes to deletion when employing a dual nuclease strategy including scarring (multifocal indels), inversions, and duplications[47,48,60,68,138]. Furthermore, the heterogeneous population of cells resulting from a dual nuclease strategy would be suboptimal for clinical translation.

Functional footprinting-informed targeting by ZFNs/TALENs within the *BCL11A* enhancer and comprehensive functional mapping of the *BCL11A* enhancer by CRISPR/Cas9-mediated saturating mutagenesis has revealed an “Achilles heel” to the *BCL11A* enhancer within the +58 DHS[147,217]. Disruption of this minimal functional sequence at the core of the DHS +58 by CRISPR/Cas9 or ZFNs/TALENs resulted in γ-globin induction comparable to targeting coding sequence in $CD34^+$ HSPCs subject to erythroid differentiation conditions[147,217]. The core region has been fine-mapped to an approximately 20 bp region including a GATA1 binding motif which appears to be essential for *BCL11A* expression and subsequent HbF repression[147,217]. As previously discussed, the erythroid specificity of the regulatory element would not require erythroid specific expression of the genome editing components, as would be necessary with a *BCL11A* coding sequence targeting approach. Taken together, targeting of the *BCL11A* enhancer disruption at the functional core of +58 DHS in autologous $CD34^+$ HSPCs followed by bone marrow transplantation represents a promising therapeutic strategy to induce HbF expression in patients with the β-globin disorders (Figure 5.2).

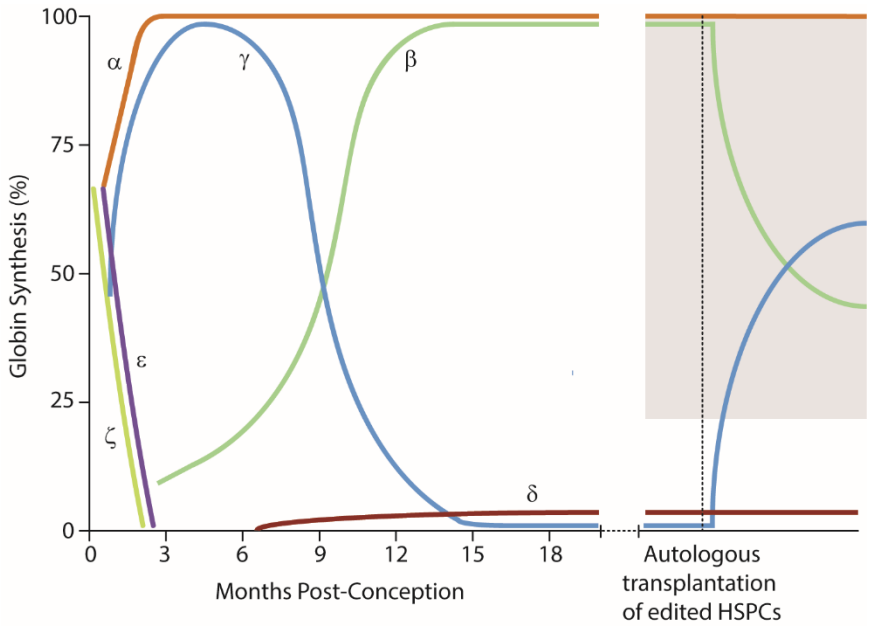


Figure 5.2: Reversal of hemoglobin switching to induce therapeutic levels of HbF. Reversal of hemoglobin switching can be accomplished through autologous bone marrow transplantation of genome edited CD34^{+} HSPCs. The gray region indicates the hypothesized levels of HbF required for clinical benefit.

LRF/ZBTB7A gene disruption

Another transcription factor LRF/ZBTB7A (also referred to as Pokemon) has more recently been recognized as a major repressor of γ-globin[218]. LRF-knockout mice exhibit elevated levels of the embryonic globin *Hbb-βh1* with normal levels of *Hbb-y*. This contrasts from *Bcl11a*-null mice that exhibit elevation of both embryonic globins, *Hbb-y* > *Hbb-βh1*[135,136]. *Zbtb7a*$^{-/-}$ mice are embryonic lethal due to anemia, whereas conditional knockout of *Zbtb7a* in adult mice leads to inefficient terminal erythropoiesis resulting in a mild macrocytic anemia[219]. CRISPR/Cas9-mediated knockout of LRF in an erythroid cell line resulted in dramatic upregulation of γ-globin. LRF/BCL11A double knockout in this system was near additive and resulted in HbF of >90%, suggesting LRF's role in γ-globin regulation is partially independent of BCL11A. Subsequent analysis demonstrated a mild delay in erythroid differentiation upon knockdown of LRF in primary human CD34^{+} HSPCs differentiated down the erythroid lineage with a corresponding induction of γ-globin[218]. While the effect of LRF loss on γ-globin is striking, the role of LRF in cell

fate decisions in multiple hematopoietic lineages and its requirement for terminal erythropoiesis may limit its therapeutic potential[220].

Reducing chain imbalance for β-thalassemias

The physiologic hallmark of β-thalassemia is globin chain imbalance, such that deficiency of β chains leads to precipitation of unstable, free α chains, membrane damage, hemolysis and ineffective erythropoiesis[20,221–223]. α-thalassemia serves as a genetic modifier of β-thalassemia, as chain imbalance is reduced[223]. This is supported by a milder disease course in patients with the co-inheritance of α-thalassemia and β-thalassemia[223–225]. In principle, therefore, α-globin genes or their regulatory elements constitute potential targets for genome editing. Targeting an α-globin gene itself could result in a heterogeneous population of cells including those null for α-globin which might not support erythropoiesis. This approach could become a viable option if technological advancements allow for the precise control of the number of α-globin null alleles generated. However, with present editing tools, the inability to control the number of α-globin null alleles created through targeting α-globin coding sequences makes targeting known regulatory elements of α-globin or RNAi-mediated knockdown of α-globin more attractive alternatives in this context.

Other potential therapeutic targets

Transcription factors KLF1 and MYB have previously been considered potential targets for HbF reactivation, but are not attractive due to their broad roles in cell proliferation and cellular development[144,197]. Other genes such as EHMT1/EHMT2 and the LIN28B pathway have been implicated in the regulation of γ-globin; however, the selectivity of these targets and roles in hematopoiesis need further investigation[226–228].

THE PLATFORM: AUTOLOGOUS BONE MARROW TRANSPLANTATION OF GENOME EDITED CELLS

Significant obstacles to wider use of bone marrow transplantation for cure of patients with β-globin disorders are the availability of compatible donors and risk of GVH disease. Donor availability is particularly severe for SCD patients. The most persuasive rationale for therapeutic genome editing of β-globin disorders rests with the use of autologous $CD34^+$ HSPCs as the cellular target. Through use of the patient's own cells for therapy, donor availability and GVH disease are avoided. As with more "conventional" somatic gene therapy with modified viruses, delivery of the requisite editing components to the target cells is the principal hurdle to be overcome in achieving clinical success[229]. Delivery of therapeutic genes to $CD34^+$ HSPCs has been accomplished with integrating and non-integrating viral vectors (such as lentiviral, adenovirus, and AAV vectors), as well as physical methods (e.g. electroporation)[178,230–232]. The optimal method for gene editing is currently unknown but likely related to the specific technology employed. High efficiency delivery at clinical scale, roughly > 10^8 $CD34^+$ HSPCs cells, presents a practical challenge. However, recent studies have taken promising steps forward with electroporation of mRNA to $CD34^+$ HSPCs at clinical scale (>1 x 10^8 cells)[217,233]. Robust cellular delivery is required for clinical translation of any envisioned therapeutic genome editing approaches.

Genome editing is generally more difficult in primary cells as compared to immortal cell lines for reasons that are not entirely well understood, but may reflect inefficient delivery, diminished promoter activity of constructs, interferon responses, exonuclease activity, and host mechanisms of DNA repair[229]. Electroporation of ZFNs, TALENs, and CRISPRs as DNA, RNA, and/or protein is an efficient delivery strategy to $CD34^+$ HSPCs in a laboratory setting[178,182,229,234,235]. For the CRISPR/Cas9 system, mRNA or ribonucleoprotein (RNP) electroporation may obviate toxicity associated with DNA delivery, as well as yield higher rates of editing in cell lines and $CD34^+$ HSPCs[187,229]. The identification of novel Cas9 proteins isolated

from diverse prokaryotes or other Cas9-like nucleases that are smaller than the widely used *Streptococcus pyogenes*-derived Cas9 may facilitate delivery efficiency particularly for viral vectors[53]. In addition, chemical modification of sgRNA's enhances editing efficiency in primary hematopoietic cells and $CD34^+$ HSPCs[229].

STEPS TO CLINICAL TRANSLATION

While clinical translation of therapeutic genome editing for the β-hemoglobinopathies is appealing, several steps must be taken before the vision can become a reality. (1) Target selection, (2) delivery of editing reagents to HSCs, and (3) empirical testing of off-target potential must all be addressed and optimized. Targets that may be chosen for clinical development are summarized in Figure 5.1 and Table 5.1. (1) Strategies that rely on NHEJ are likely to be the first attempted using current technologies due to the dominance of NHEJ in quiescent HSCs and overall high efficiency of NHEJ as compared to HDR. At present, modification of the core *BCL11A* enhancer sequences of the +58 DHS of the *BCL11A* gene by NHEJ appears quite favorable in terms of potency of an effect on HbF expression and sparing of consequences for non-erythroid lineages. (2) Transient delivery (electroporation or non-integrating viral vectors) represents a safer alternative to stable integration of genome editing components due to reduced risk of insertional mutagenesis and risk of off-target cleavage, as well as freedom from the uncertainty of long-term expression of genome editing tools in $CD34^+$ HSPCs. Transient delivery also necessitates high levels of on-target editing within a shorter window of time prior to loss of the genome editing components through cell division. One possibility would be to enrich for edited cells prior to bone marrow transplantation of autologous cells[60], which could be further enhanced by strategies to expand HSCs *ex vivo*[189–191]. (3) Off target cleavages represent a legitimate concern for therapeutic genome editing. Newly developed techniques allow for unbiased genome-wide identification of off target mutagenesis[236,237]. Various methods have been reported that aim to enhance on-target versus

off-target specificity. These include use of Cas9 nickase, truncated guides, dimeric RNA-guided FokI nucleases, and rationally engineered enhanced specificity Cas9[51,52,68,95,238,239]. In addition, alternative RNA-directed nucleases (Cpf1) or modified Cas9-derivatives with reduced off-target cleavage potential appear to be steps toward "clean" editing reagents. It will be necessary to empirically test the optimized editing reagents for off-target cleavage potential and assess the associated risk of inappropriate DSBs within the genome. As methods to predict and detect off-target cleavages continue to improve, it may become possible to screen autologous genome edited cells prior to bone marrow transplantation for possible pathogenic off-target mutations. “CD34$^+$ humanized” mice, NOD-SCID-Gamma mice with bone marrow engrafted human CD34$^+$ HSPCs, can be used to evaluate safety of genome editing tools as these models can demonstrate multilineage reconstitution, self-renewal, and the ability to monitor leukemogenesis. However, due to the inability to model all human hematopoietic lineages, notably the erythroid lineage, and general limitations of chimera mouse models, humanized mice have limitations in assessing safety of genome editing treatments *in vivo*.

One additional challenge for clinical development is harvesting CD34$^+$ HSPCs for autologous stem cell transplantation from patients with β-globin disorders. Sufficient numbers of CD34$^+$ HSPCs for bone marrow transplant can be harvested from two sources, peripheral blood or bone marrow. Harvest of CD34$^+$ HSPCs from peripheral blood is preferred due its minimal invasiveness and higher yield of CD34$^+$ HSPCs following mobilization by granulocyte colony-stimulating factor (G-CSF)[240]. Use of G-CSF as a mobilizing agent is generally well-tolerated for healthy adults and cancer patients. However, there are significant risks of G-CSF administration for patients with β-globin disorders. SCD patients have significant risk of vaso-occlusive events, acute chest syndrome, multi-organ system failure, and death[241] whereas β-thalassemia patients are susceptible to splenic rupture, hyperleukocytosis, and thrombosis[240]. Plerixafor is an alternative mobilizing agent that may provide a safer option to G-CSF[240,242,243]. While the effect of plerixafor in SCD patients requires investigation, it has been shown to be safe and effective in

both splenectomized and non-splenectomized β-thalassemia patients. In contrast, while G-CSF was well-tolerated in non-splenectomized patients, it resulted in hyperleukocytosis and lower yield of $CD34^+$ HSPCs as compared to plerixafor in splenectomized β-thalassemia patients[242]. Combination of plerixafor with a reduced dose of G-CSF to avoid adverse effects has been shown to be superior to either agent alone[240,243]. Therefore, plerixafor or combination G-CSF/plerixafor mobilization may provide a safe avenue for peripheral blood $CD34^+$ HSPC harvests for β-thalassemia patients. Until acceptable protocols for mobilization of $CD34^+$ HSPCs are established for SCD patients, traditional bone marrow harvesting may be required. It may be advisable as well to test *ex vivo* editing efficiencies and maintenance of modified cells upon transfer into suitable immunodeficient mice for $CD34^+$ HSPCs obtained by different methods to ensure optimization for clinical use.

CONCLUSIONS

The technological advances in genome manipulation are breathtaking in terms of the speed with which they have been reported in the past several years. The potential of genome editing approaches for clinical benefit in the β-globin disorders is immense. Besides the choice of the editing platform and its delivery to repopulating cells within $CD34^+$ HSPC harvests, a major factor in considering application to these conditions is the target sequences to be modified. If the goal is precise gene correction, the desired sequence alteration is clear. This strategy relies on HDR, and at the moment must await improved protocols for HDR in *bona fide* repopulating cells for clinical experimentation.

Reactivation of HbF is an attractive approach, as it might be "one size fits all" in principle, suitable for both SCD and the β-thalassemias. The precise levels of pancellular HbF necessary for clinical benefit remain elusive, but is hypothesized to be ≥20% for SCD and likely somewhat higher in β-thalassemia (Figure 5.2). Due to the inability to accurately model HbF experimentally, it may be difficult to assess the minimal threshold for clinical benefit in a

laboratory setting. Further, it is unlikely that HSCs undergoing therapeutic genome editing based on the strategies outlined above will have a selective advantage *in vivo*. However, results from mixed chimerism allogeneic transplant demonstrate that low levels of chimerism can produce clinical benefit due to the survival advantage of normal RBCs[168].

Given the current state of genome editing technologies, HbF induction mediated by NHEJ repair may provide a long-sought “silver bullet” for therapy. As such, harnessing the power of genome editing tools may finally allow for therapeutic exploitation of the deep understanding of the genetics of hemoglobin and lead to a genome editing-based therapeutic option for the β-hemoglobinopathies in the near future.

Appendix: Supplemental Material

CHAPTER 3 SUPPLEMENTAL MATERIAL

CHAPTER 3 SUPPLEMENTAL TEXT

Evaluation of human *BCL11A* enhancer in immortalized human erythroid precursors

We utilized HUDEP-2 cells, an immortalized human $CD34^+$ hematopoietic stem and progenitor cell (HSPC)-derived erythroid precursor cell line that expresses BCL11A and predominantly β- rather than γ-globin[122]. We used the CRISPR-Cas9 nuclease system to generate a clone of HUDEP-2 cells null for *BCL11A* by deleting coding sequences (Figure 3.1a). These cells demonstrated elevated levels of γ-globin mRNA and HbF protein, consistent with the functional requirement of BCL11A for HbF repression (Figure 3.1b, c). Deletion of the 12-kb *BCL11A* composite enhancer with a pair of sgRNAs resulted in near complete loss of BCL11A expression and induction of γ-globin and HbF protein to similar levels as cells with *BCL11A* knockout (Figure 3.1a-c), analogous to the requirement of the orthologous mouse composite enhancer for erythroid BCL11A expression[67]. Introduction of Cas9 and an individual sgRNA targeting *BCL11A* exon-2 to HUDEP-2 cells as well as to primary human erythroid precursors produced cells with elevated HbF expression, indicating loss of BCL11A function and resultant derepression of BCL11A's target γ-globin (Figure 3.2g, h; Supplemental Figure 3.2e). The level of BCL11A transcript was unaffected in these cells (Figure 3.2f), suggesting that *BCL11A* transcripts with protein truncating variants (due to frameshift or nonsense mutations) escape nonsense-mediated decay.

Pooled CRISPR enhancer saturating mutagenesis screen in the HUDEP-2 cells

We designed all possible sgRNAs within the human *BCL11A* composite enhancer DHSs (Figure 3.1d-f) as restricted only by the presence of the SpCas9 NGG protospacer adjacent motif

(PAM)[47,93]. The NGG PAM restricted sgRNAs had a median adjacent genomic cleavage distance of 4 bp and 90th percentile of 18 bp (Figure 3.1f), which suggested that this strategy could approach saturation mutagenesis *in situ* given the expected indel spectrum produced by each sgRNA to include frequent deletions up to 10 bp from the cleavage position[47,48,60]. The production of indels within an enhancer sequence could alter gene regulation by a variety of non-mutually exclusive mechanisms including: (1) disruption of transcription factor (TF) binding motifs, (2) creation of TF binding motifs, (3) changes in spacing of TF binding motifs without direct disruption of the binding sequences themselves, and (4) effects independent of TF binding motifs.

NAG may act as an alternate PAM for SpCas9, albeit with lower efficiency[93]. We also designed sgRNAs restricted by the NAG PAM (Supplemental Figure 3.2a, b). The NAG PAM restricted sgRNAs had a median adjacent genomic cleavage distance of 5 bp and 90th percentile of 15 bp (Supplemental Figure 3.2b).

We included 120 nontargeting sgRNAs as negative controls as well as 88 sgRNAs tiling exon-2 of *BCL11A* as positive controls. The total library included 1,338 sgRNAs (Supplemental Figure 3.2c). We synthesized oligonucleotides for the NGG and NAG restricted and nontargeting sgRNAs on a microarray and cloned the sgRNAs as a pool to a lentiviral vector[84]. Deep sequencing of the lentiviral plasmid library (including both NGG and NAG sgRNAs) demonstrated that 1,337 of 1,338 sgRNAs (99.9%) were successfully cloned. The representation of sgRNAs within the library showed a relatively narrow distribution, with a median of 718 and the 10% and 90% percentile ranging from 337 to 1,205 normalized reads (Supplemental Figure 3.2d). Since the number of viral integrants per cell follows a Poisson distribution, maximizing single integrants can be achieved by decreasing multiplicity of infection[244–246]. Therefore, the basic experimental schema was to transduce cells with the lentiviral library at low multiplicity such that nearly all selected cells contained a single integrant (Figure 3.1d). We transduced HUDEP-2 cells stably expressing SpCas9 with the pooled library

of *BCL11A* enhancer targeting sgRNAs. We initially expanded the cells for one week, and subsequently transferred them to erythroid differentiation conditions, for a total of two weeks of culture. Then we performed intracellular staining for HbF. Fluorescence activated cell sorting (FACS) was employed to isolate HbF-high and HbF-low pools (consistent with low and high BCL11A activity respectively; Figure 3.1d; Supplemental Figure 3.2e, g). We enumerated the representation of the library in each pool by deep sequencing. The HbF enrichment score of each sgRNA was calculated as the $\log_2$-ratio of normalized reads in the HbF-high compared to HbF-low pools. We compared the HbF enrichment of the 120 nontargeting negative control sgRNAs and 88 coding sequence targeted positive controls for both NGG and NAG PAM restricted sgRNAs. We observed equivalent representation of the nontargeting sgRNAs in the HbF-high and HbF-low pools but highly significant enrichment of the NGG sgRNAs targeting exon-2 of *BCL11A* in the HbF-high pool, consistent with a reduction of BCL11A activity (Supplemental Figure 3.2h). One nontargeting sgRNA (#0548) had an HbF enrichment score of 0.803, while the remaining 119/120 nontargeting sgRNAs (99.2%) showed enrichment scores below 0.259. In contrast 40/48 sgRNAs targeting BCL11A exon 2 (83.3%) showed enrichment scores above 0.259. These results suggest that the large majority of sgRNAs in the library were competent to produce indels. However, exon-2 targeting sgRNAs with NAG PAM restriction did not show significant HbF enrichment suggesting inefficient indel production (Supplemental Figure 3.2h). Therefore the NAG restricted sgRNAs were excluded from further analyses.

We compared the representation of sgRNAs in the initial plasmid pool to the representation of sgRNAs in the cells at the end of *in vitro* culture. While the majority of the library maintained neutral representation throughout the experiment, we observed a fraction of sgRNAs that were depleted, mainly among the h+62 sgRNAs (Supplemental Figure 3.2i). We observed that these dropout sgRNAs overlapped with repetitive elements within the genome, in particular to a SINE AluSq element that appears in the genome nearly 100,000 times[247] (Supplemental Figure 3.2k). Initial design of sgRNAs did not include prediction of off-target

cleavage to maximize the resolution of target mutagenesis. We removed from subsequent analysis the 35 of 582 (6.0%) NGG PAM sgRNAs with cellular dropout from the plasmid pool greater than 8-fold, since these dropouts indicated apparent *BCL11A*-independent effects of genomic disruption (Supplemental Figure 3.2i, k).

The majority of enhancer targeting sgRNAs showed no significant enrichment or depletion from the HbF-high pool (Supplemental Figure 3.2j). We observed a number of sgRNAs with HbF enrichment at each of the DHSs as well as some with HbF depletion at h+55 (Supplemental Figure 3.2j). We observed discrete sets of colocalizing sgRNAs with elevated HbF enrichment, with a particularly robust cluster at h+58 (Figure 3.2a). Our screen data appear consistent with the effects of indels as produced by individual sgRNAs rather than combinations of sgRNAs for several reasons: (1) negative control sgRNAs did not show evidence of impact on BCL11A expression, arguing against the occurrence of prevalent passenger effects (Figure 3.2a); (2) positive control sgRNAs targeting BCL11A coding sequences (in which the expected outcome of individual sgRNAs would be frameshifted null alleles) show that the great majority produce the expected effect of HbF derepression (Figure 3.2a); (3) the enhancer targeting sgRNAs largely follow a null distribution of effect sizes with the exception of a few outliers suggesting most of the enhancer targeting sgRNAs had no impact on BCL11A expression again arguing against prevalent passenger effects (Supplemental Figure 3.2j, 3.7b); and (4) the enhancer targeting outliers with positive enrichment map to colocalizing discrete genomic sequences (Figure 3.2a, 3.3, 3.5a) which would appear unlikely if the impact on BCL11A expression were due to multifocal indels, large deletions, inversions, rearrangements, or translocations (the expected outcomes of more than one genomic cleavage) but likely if the impact were due to individual indels (the expected outcomes of single genomic cleavages).

A limitation of our primary pooled screen approach is that the measurements of the enrichment of sgRNAs come from a pooled population of cells that were transduced by the entire library. Therefore we endeavored to prospectively validate the results of individual

sgRNAs as identified by the screen. We observed a strong correlation between the HbF enrichment score from the screen and the fraction of HbF^+ cells in arrayed format, testing 24 sgRNAs with enrichment scores ranging from the highest to the lowest in the screen, and representing sgRNAs from all 5 mapping categories ($r = 0.816$, $p < 0.0001$; Supplemental Figure 3.3a, b). These results demonstrate that a single enhancer-targeting sgRNA may mediate robust HbF induction.

Common genetic variation in functional enhancer cores

The h+62 Active region contains only one common SNP (MAF>1%), the variant rs1427407, which was previously identified by fine-mapping as the most highly trait-associated SNP[67] (Figure 3.3c; Supplemental Figure 3.5c). The high-HbF T-allele is disruptive of an apparent half E-box/GATA composite motif ($P = 9.74 \times 10^{-4}$ for T-allele, $P = 1.69 \times 10^{-4}$ for G-allele, though neither met our predefined threshold for significance of $P < 10^{-4}$) and associated with reduced GATA1 and TAL1 occupancy in primary human erythroid chromatin as previously described[67]. Multiple sgRNAs with cleavages mapping directly to the motif demonstrated positive enrichment scores (Supplemental Figure 3.5c). Of note, there was a gap of 88 nucleotides between sgRNA cleavages at the core of the Active region due to lack of NGG PAM motifs. Despite this uncommon limitation of functional resolution by SpCas9 and NGG PAM restricted sgRNAs (Figure 3.1f), the HMM model was still able to identify the region. Substantial interspecies conservation as evaluated by both PhyloP and PhastCons (which model individual nucleotide and multibase element conservation, respectively) was observed at this h+62 Active state region as compared to flanking regions (Figure 3.3c, Supplemental Figure 3.5c).

DHS h+55 encompasses the SNP rs7606173, which along with rs1427407 defines the most highly trait-associated haplotype (Figure 3.3a; Supplemental Figure 3.5a). Previous fine-mapping was unable to find additional SNPs at *BCL11A* with predictive power for the trait association beyond the rs1427407–rs7606173 haplotype based on conditional or rare-variant

analyses. No common SNPs were found directly within the Active or Repressive state regions of h+55, however rs7606173 resides merely 3 bp from the Repressive region and 34 bp from the Active region (Supplemental Figure 3.5a). The next closest common SNP to an Active or Repressive state within h+55 is rs62142646, which is 739 bp from an Active state. The major, ancestral G allele at rs7606173 is associated with high-HbF. The HUDEP-2 cells used in this screen are homozygous for this G variant. Given a model in which high-HbF trait is due to disruption of TF binding sequences at the *BCL11A* enhancer, sgRNA-mediated disruption of the high-HbF rs7606173-G allele might not be expected to lead to further functional impact. We did observe six motifs predicted ($P < 10^{-4}$) to be differentially impacted by the rs7606173 genotype (Supplemental Figure 3.5a). The top-scoring sgRNAs in h+55 cluster 56-58 bp from rs7606173, at a site with a predicted TAL1::GATA1 motif ($P < 10^{-4}$). This sequence element possesses high vertebrate conservation (Supplemental Figure 3.5a). The entire region encompassing the Active/Repressive h+55 states appears to have elevated sequence conservation as compared to flanking sequences (Figure 3.3a).

The only common SNP within the h+58 Active region is rs6738440 just at the edge of the Active state region (chr2:60722241), 118 to 160 bp from the cluster of top-scoring sgRNAs (chr2:60722359-60722401; Figure 3.4; Extended Figure 3.5b); the next closest common SNP was rs62142615 (chr2:60722120), 119 bp away. Neither sgRNAs with significant adjacent enrichment nor overlying genome-scale significant motifs with either the major A- or minor G-allele were observed at rs6738440. Previous conditional analysis of the rs1427407-rs7606173 haplotype was unable to demonstrate residual significant trait association for this variant[67].

Enhancers paradoxically may demonstrate both evolutionary conservation and heightened turnover. Common trait-associated enhancer variation suggests the frequent occurrence of intraspecies polymorphic sequences sufficient to modulate enhancer function and thereby produce novel phenotypes. Per above, we previously described that the trait-associated enhancer haplotype at *BCL11A* is defined by two SNPs[67]. Our pooled CRISPR screening

revealed that each of these SNPs reside near functional enhancer states consistent with their roles as causal variants. The most potent enhancer region, within h+58, has no trait-associated variants near its functional core. This example demonstrates how fine-mapping GWAS associations to individual SNPs can substantially underestimate the biologic importance of the underlying elements to the associated trait.

Pooled CRISPR enhancer saturating mutagenesis screen in mouse erythroid εy:mCherry reporter cells

We generated a MEL cell reporter line with the mCherry fluorescent reporter knocked-in to the embryonic globin *Hbb-y* locus (Supplemental Figure 3.6c). Introduction of Cas9 and sgRNA targeting *Bcl11a* exon-2 resulted in the appearance of cells with elevated εy:mCherry expression, indicating derepression of the BCL11A target εy-globin (Supplemental Figure 3.6h).

The mouse sgRNA library was comprised of both NGG and NAG PAM restricted sgRNAs. The library included sgRNA sets tiling the DHS m+55, m+58, and m+62 orthologs, as well as 120 nontargeting negative controls and 91 *Bcl11a* exon-2 targeting positive controls (Supplemental Figure 3.6d, g). Similar to the human enhancer screen, the sgRNAs were distributed throughout the target sites, with a median distance to adjacent cleavage site of 4 bp and 90% of adjacent cleavage sites falling within 18 bp for NGG PAM restricted sgRNAs (Supplemental Figure 3.6e). We successfully cloned into lentiviral plasmids all 1271 members of the library with a relatively narrow distribution of representation (median 735, 10%ile 393, 90%ile 1240 normalized reads; Supplemental Figure 3.6f).

Following transduction at low multiplicity by the lentiviral library, and *in vitro* culture for two weeks, cells were sorted into high- and low-εy:mCherry pools (Supplemental Figure 3.6i). Deep sequencing was performed of the genomic DNA to evaluate the representation of sgRNA libraries in the pools. The nontargeting negative control sgRNAs were evenly represented in the high- as compared to low-εy:mCherry pools whereas the positive control *Bcl11a* exon-2

targeting sgRNAs with NGG PAM were significantly overrepresented in the εy:mCherry-high pool (Supplemental Figure 3.6j). Although there was slight enrichment that reached statistical significance, the NAG PAM restricted sgRNAs showed substantially reduced overrepresentation relative to the potent NGG restricted sgRNAs, so further analysis was restricted to the NGG PAM restricted sgRNAs (Supplemental Figure 3.6j).

We analyzed the representation of the library in cells that had completed two weeks of *in vitro* culture (sum of the high- and low-εy:mCherry pools) as compared to the initial lentiviral plasmid pool. The large majority of sgRNAs showed equivalent representation in the initial plasmid pool and as integrants in cells at the completion of the experiment (Supplemental Figure 3.7a). A small number of sgRNAs (n=8) showed substantial cellular dropout $>2^{-3}$ and were removed from subsequent enrichment analysis. Similar to the human screen, these mapped to repetitive elements (Supplemental Figure 3.7c).

We determined εy enrichment score as the $\log_2$-ratio between representation in the εy:mCherry-high as compared to εy:mCherry-low pools (Figure 3.5a; Supplemental Figure 3.7a). We noted almost all exon-2 targeting sgRNAs demonstrated both positive εy enrichment scores and negative cellular dropout scores with high correlation (Figure 3.5a; Supplemental Figure 3.7a, c, d).

The majority of enhancer targeting sgRNAs showed no significant εy enrichment (Supplemental Figure 3.7b). We detected sgRNAs with both modest enrichment and depletion from the εy:mCherry-high pool at the m+55 ortholog, similar to h+55. We detected a set of sgRNAs with marked εy enrichment at the m+62 ortholog, exceeding the potency of those enriching at h+62. At the m+58 ortholog we did not observe any evidence of εy enriching or depleting sgRNAs (Supplemental Figure 3.7b).

We applied the same HMM model to infer Active, Repressive, and Neutral states at the mouse BCL11A enhancer orthologs (Supplemental Figure 3.4a, 3.8a-c). We identified an Active state at the m+62 ortholog and Active and Repressive states at the m+55 ortholog. Only the

Neutral state was identified at the m+58 ortholog. The regions of the m+55 and m+62 DHSs with peak DNase I sensitivity were inferred as possessing Active states (Supplemental Figure 3.8a-c). We analyzed 108 clones in which the entire composite enhancer was first monoallelically deleted and subsequent hemizygous mutations were produced by sgRNAs targeting the m+62 ortholog on the remaining allele. We measured BCL11A expression by RT-qPCR in each of these 108 clones normalized to 25 control clones not exposed to m+62 targeting sgRNAs. This clonal analysis identified a core region of the m+62 ortholog containing functional sequences required for BCL11A expression and embryonic εγ-globin repression (Supplemental Figure 3.8d, 9). The region is rich with TF-binding motifs, particularly those of key factors involved in erythropoiesis and globin gene regulation, including Gata1, Klf1, and Myb (Supplemental Figure 3.9). Of note, despite the presence of relatively high vertebrate conservation throughout the m+62 and h+62 Active state regions (Figure 3.4c, Supplemental Figure 3.5c, 3.8c, 3.9b), the impact of the m+62 ortholog on BCL11A and globin gene regulation greatly exceeded that of h+62 (Figure 3.2a, c-e, 3.5a-c; Supplemental Figure 3.8c, d, 3.9).

Human and mouse sequence and functional conservation

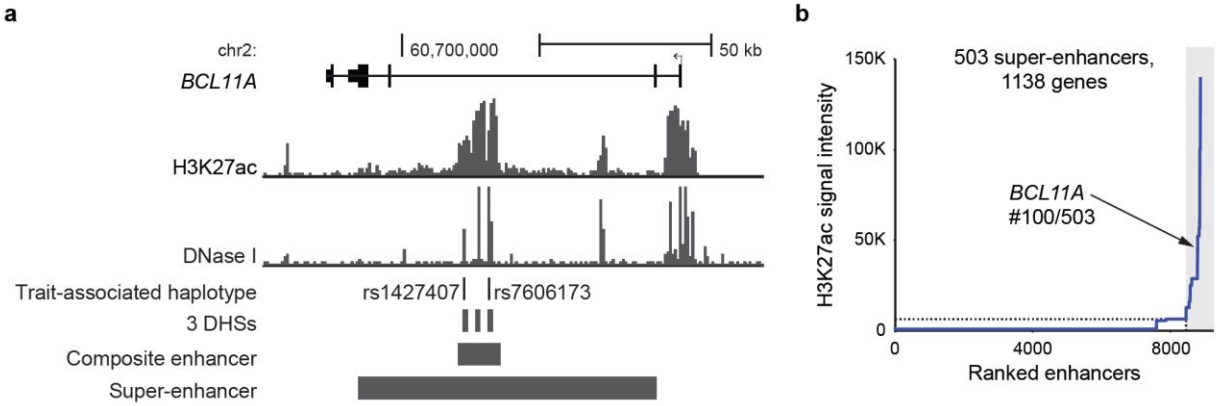
Sequence homology is detectable at an approximately similar distal intron-2 position with respect to the TSS for each of the mouse sequences homologous to the three human DHSs: h+55 (length 1283 bp) has 402 positions of nucleotide identity (31.3%) compared to the m+55 ortholog (length 1046 bp), h+58 (1264 bp) has 367 positions of nucleotide identity (28.6%) compared to the m+58 ortholog (length 1341 bp), and h+62 (length 1369 bp) has 281 positions of nucleotide identity (20.5%) compared to the m+62 ortholog (length 1216 bp). By comparison, of the 2508 bp in human *BCL11A* coding sequence (XL isoform), 2424 nucleotides demonstrate identity (96.7%) compared to mouse *Bcl11a* coding sequence (XL isoform).

Although enhancers are composed of TF binding motifs, the presence of motifs alone is inadequate to predict enhancers. Motif predictions can be overly sensitive, in that only a small

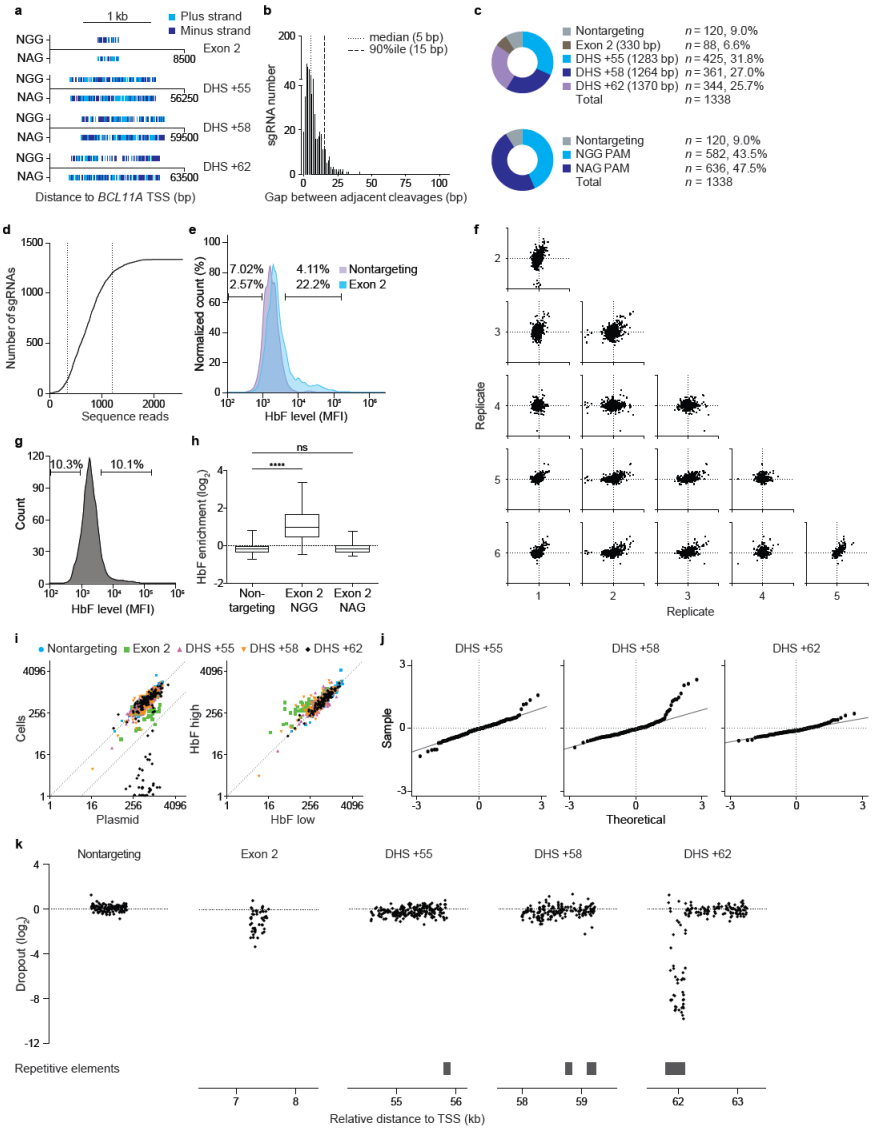
fraction of predicted motifs tend to be corroborated by ChIP-seq occupancy studies. On the other hand, motif prediction can also be insensitive; for example, a recent report highlights the importance of low-affinity motifs for achieving specificity of enhancer function[248]. Previously we showed that GATA1 occupies h+58 in primary human erythroid precursors[67]. However the orthologous m+58 region possesses neither DNase sensitivity nor functional requirement in mouse erythroid cells. Despite this divergence, the human core GATA1 motif has a similar *P*-value in the nonfunctional mouse ortholog. These results are consistent with a model in which the motif context is critically important in enhancer activity. The sequences immediately adjacent to the GATA1 motif, where both HbF-associated sgRNAs and mutations enrich, are candidates to fulfill this contextual requirement.

Recent appreciation for the wide variation in intensity of biochemical features associated with enhancer elements has led to a renewed interest in clustered enhancer elements and so-called super-enhancers. Based on H3K27ac ChIP-seq in primary human adult erythroid precursors, the composite *BCL11A* enhancer scores as a human erythroid super-enhancer (Supplemental Figure 3.1a, b). We used published H3K27ac ChIP-seq data from mouse erythroid cells and found variable results, in one dataset[128] *Bcl11a* scores as a mouse erythroid super-enhancer whereas in another dataset[127] it ranks below the super-enhancer threshold (Supplemental Figure 3.6a, b). Here we provide an example of a super-enhancer organized as a hierarchy of constituent DHSs, with some critical and others minimally required for gene expression. Even within a critical substituent DHS such as *BCL11A* h+58, there are many dispensable and only a few critical sequences. These experiments show how a super-enhancer may be vulnerable to indels produced by single DSBs.

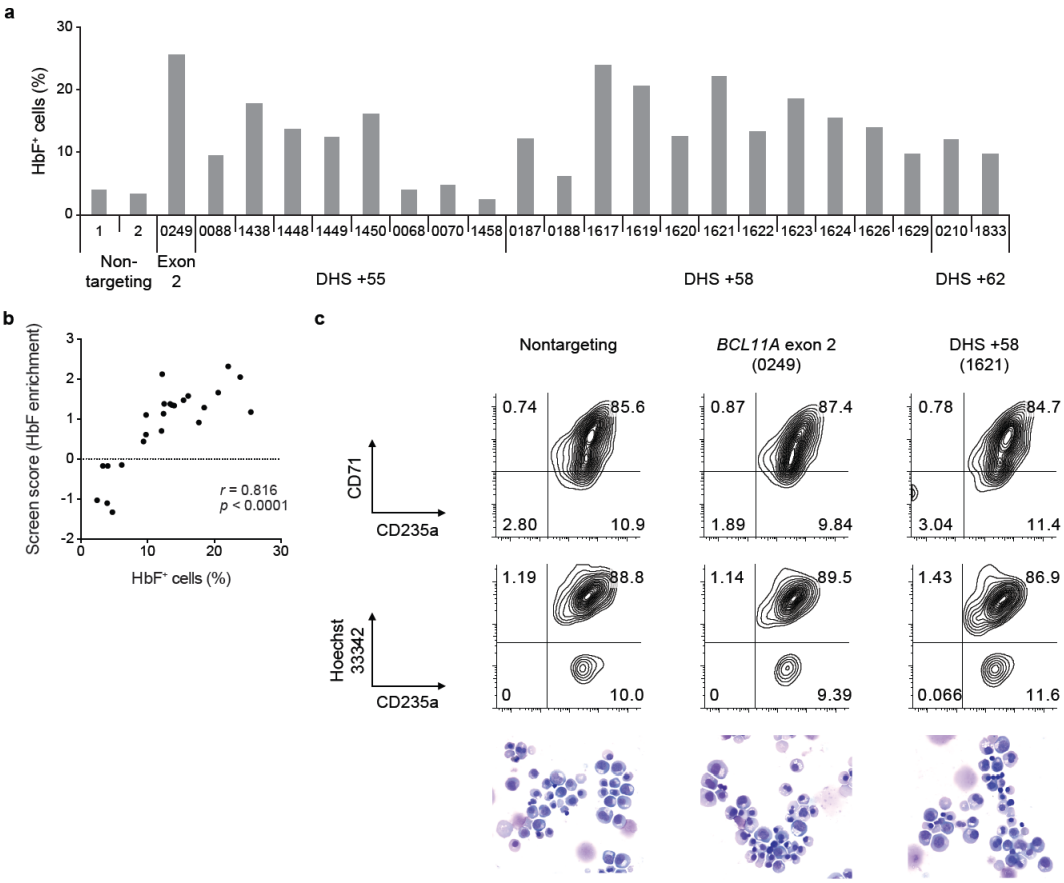
CHAPTER 3 SUPPLEMENTAL FIGURES AND TABLES



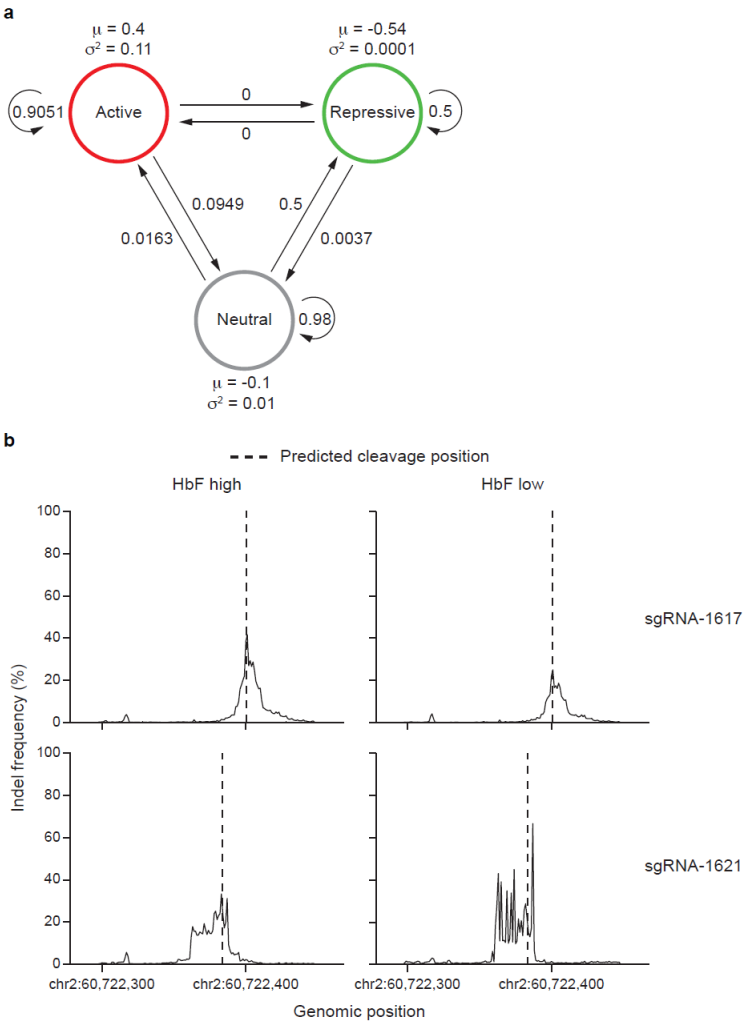
Supplemental Figure 3.1: Human *BCL11A* locus. a, Schematic of the human *BCL11A* locus (hg19, transcription from right to left) with erythroid chromatin marks and trait-associated haplotype denoted, and composite enhancer as previously defined[67]. **b**, Ranked enhancers in primary human adult erythroid precursors by H3K27ac signal intensity, with super-enhancers shaded, and super-enhancer associated genes indicated.



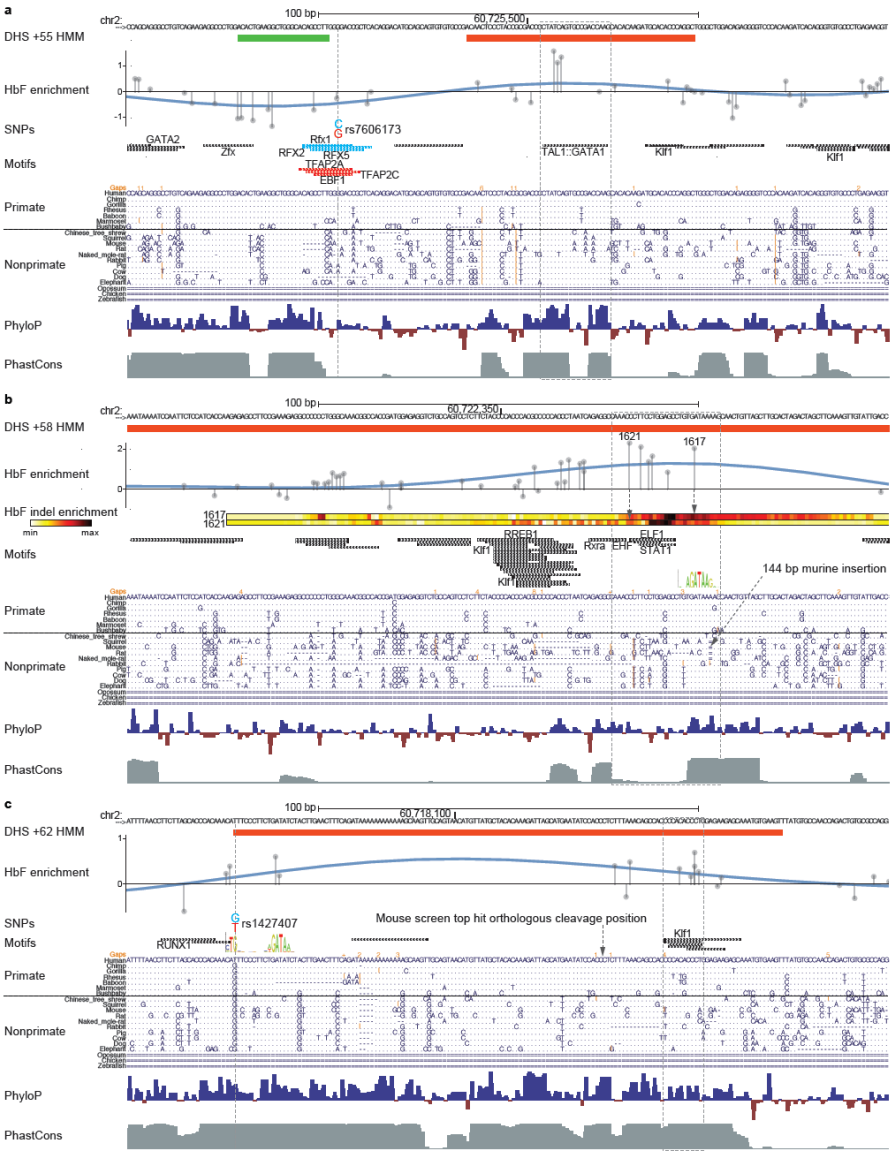
Supplemental Figure 3.2: Tiled pooled *in situ* CRISPR-Cas9 *BCL11A* enhancer screen. a, Distribution of NGG and NAG PAM sgRNAs mapped to genomic cleavage position. The vertical lines represent cleavage sites for sgRNAs mapped to plus and minus strands. **b**, Gap distance between adjacent genomic cleavage position for NAG PAM sgRNAs. **c**, Library composition by target sequence and PAM restriction. **d**. Representation of both NGG and NAG sgRNA (1,338 sgRNAs in total) within the plasmid pool by deep-sequencing. The median was 718 normalized reads and the 10th and 90th percentiles (indicated by the vertical dotted lines) ranged from 337 to 1,205 normalized reads. **e**, HbF distribution in HUDEP-2 cells transduced with Cas9 and individual sgRNAs, either nontargeting or targeting *BCL11A* exon 2. **f**, HbF enrichment scores of NGG sgRNAs in six biological replicates. **g**, Sort of library-transduced cells into HbF-high and HbF-low pools. **h**, Control sgRNA enrichment. Boxes demonstrate 25th, median, and 75th percentiles and whiskers minimum and maximum values. **** $P < 0.0001$, ns non-significant. **i**, NGG sgRNA representation in plasmid pool and cells at conclusion of experiment (*left*), and in HbF-high and HbF-low pools (*right*), with dotted lines at x=y and x=8y. **j**, Quantile-quantile plots of NGG sgRNA enrichment scores. **k**, Cellular dropout scores of NGG sgRNAs relative to genomic cleavage position and repetitive elements. Nontargeting sgRNAs pseudo-mapped with 5 bp spacing.



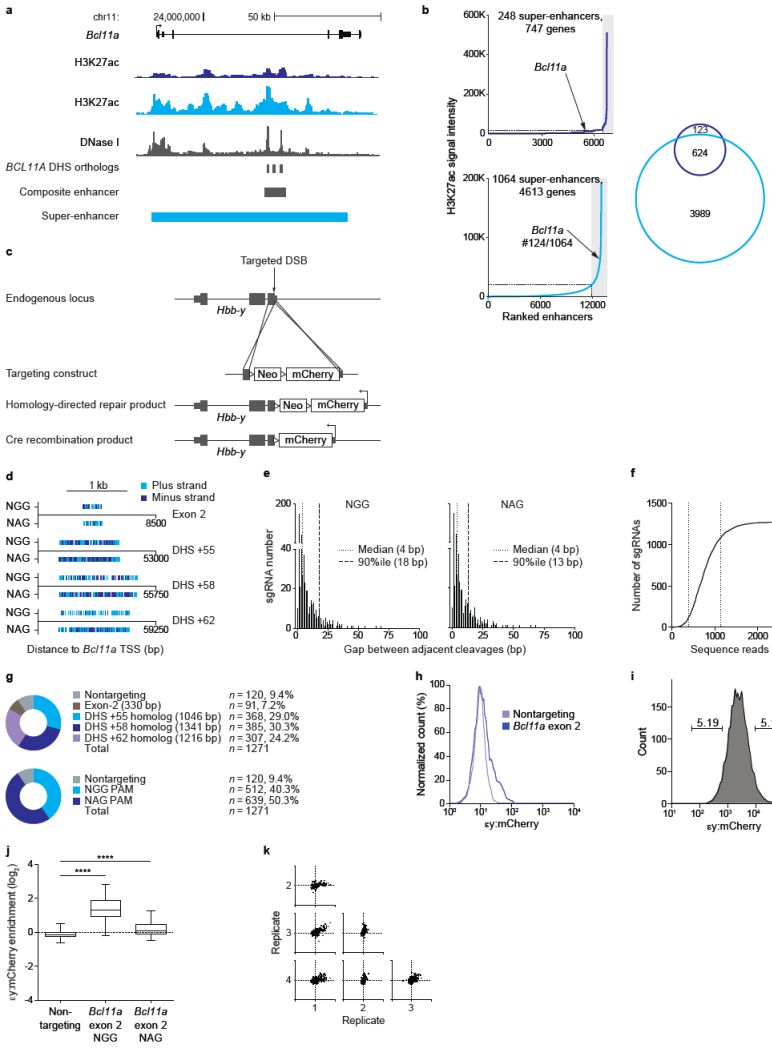
Supplemental Figure 3.3: Validation of enhancer screen. **a**, HbF^+ fraction in HUDEP-2 cells transduced in arrayed format with 24 sgRNAs from all 5 mapping categories with enrichment scores ranging from the highest to the lowest in the screen. **b**, Correlation between HbF enrichment score from pooled sgRNA screen and HbF^+ fraction by arrayed validation of individual sgRNAs in HUDEP-2 cells. **c**, Erythroid differentiation of primary human erythroid precursors evaluated by CD71 and CD235a surface markers, enucleation frequency ($CD235a^+$ $Hoescht33342^-$), and morphology by May-Grünwald-Giemsa staining.



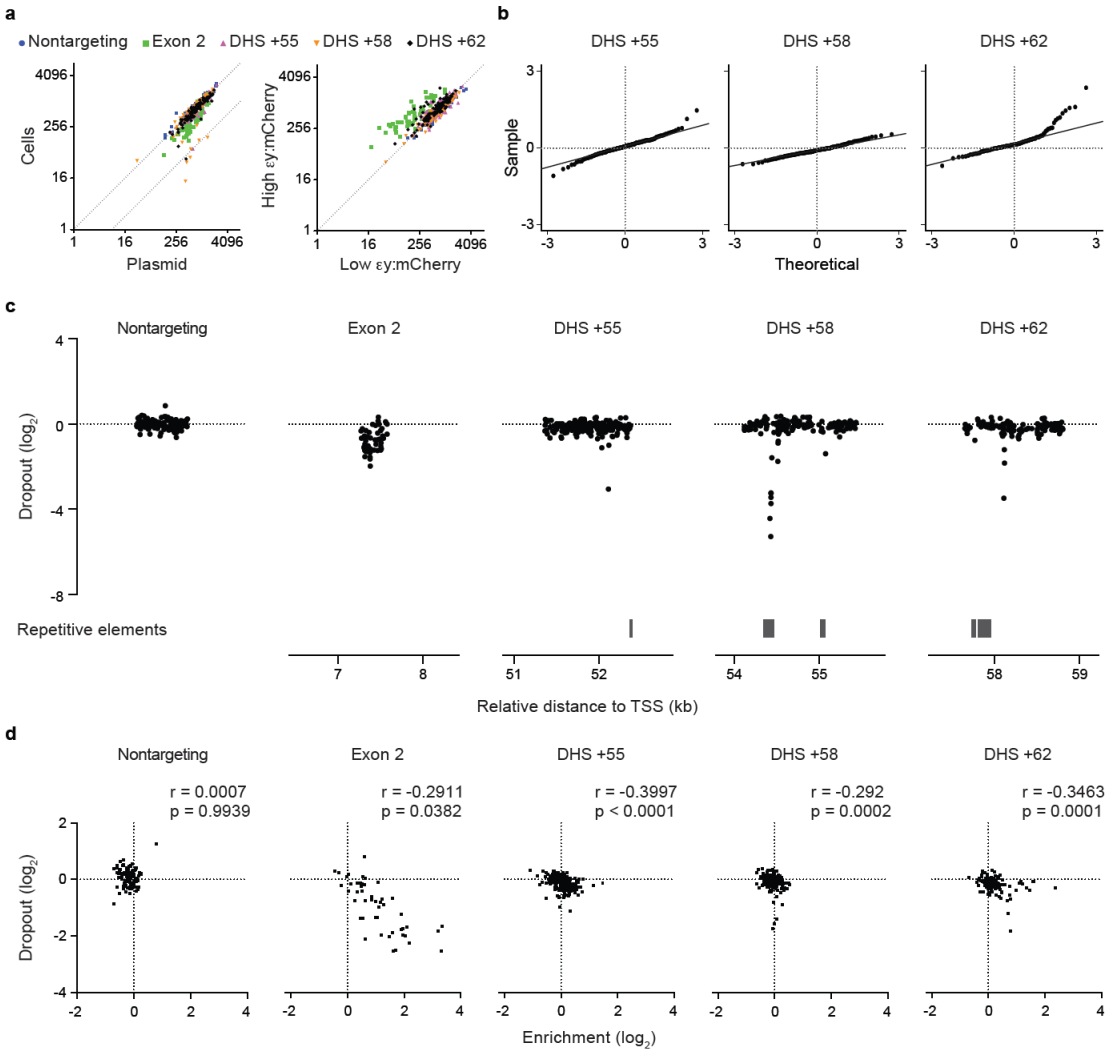
Supplemental Figure 3.4: Functional assessment of enhancer sequences. a, Topology of the Hidden Markov model (HMM) used to infer the three functional enhancer states (Active, Repressive, and Neutral). The emission probabilities of HbF enrichment scores from each state were modeled as Gaussian distributions (the values of μ and σ^2 are shown). The transition probabilities (arrows) are displayed. **b**, Frequency distribution of indels from HUDEP-2 cells exposed to Cas9 and individual sgRNAs, sorted into HbF-high and -low pools, and subjected to deep sequencing of the target site. Indels calculated on a per nucleotide basis throughout an amplicon surrounding the sgRNA-1617 and -1621 cleavage sites (dotted lines). An indel enrichment ratio was calculated by dividing normalized indel frequencies in the HbF-high pool by those in the HbF-low pool.



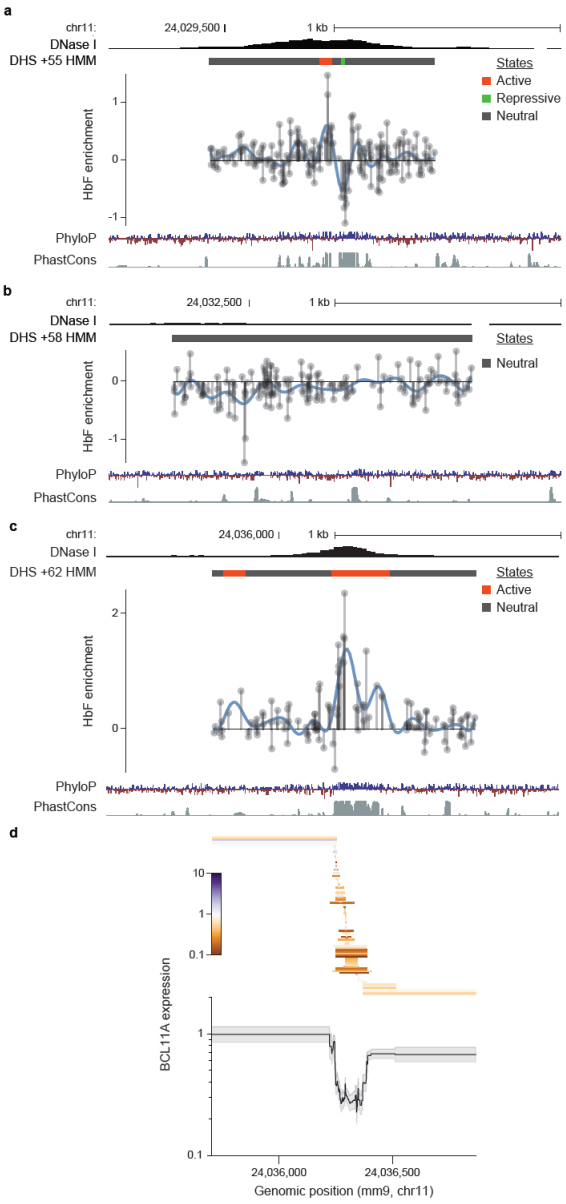
Supplemental Figure 3.5: Functional cores of the *BCL11A* enhancer. a-c, 200 bps at the functional cores of DHSs h+55, h+58, and h+62 defined by HMM states (Active red, Repressive green). HbF enrichment scores shown by gray lines and circles. HbF indel enrichment per nucleotide based on amplicon genomic sequencing of sorted cells exposed to either sgRNA-1617 (*top*) or -1621 (*bottom*). Common SNPs (MAF>1%) shown with dotted lines with HbF-low allele in blue and HbF-high allele in red; no common SNPs present at h+58 region. JASPAR motifs ($P < 10^{-4}$) depicted in black except for those with allele-specific significance depicted by allelic color. Selected motifs annotated by TF based on known erythroid-specific function or genomic position. Motif LOGOs at key positions with motif scores $P < 10^{-3}$ as described in text. Dotted boxes show regions of highest HbF enrichment score at each core with underlying predicted motifs. Orthologous sequences listed from representative primates and nonprimates of distributed phylogeny. PhyloP (scale from -4.5 to 4.88) and PhastCons (from 0 to 1) estimates of evolutionary conservation among 100 vertebrates. An arrow indicates a 144 bp insertion in the mouse genome relative to the human reference adjacent to the orthologous GATA1 motif at h+58.



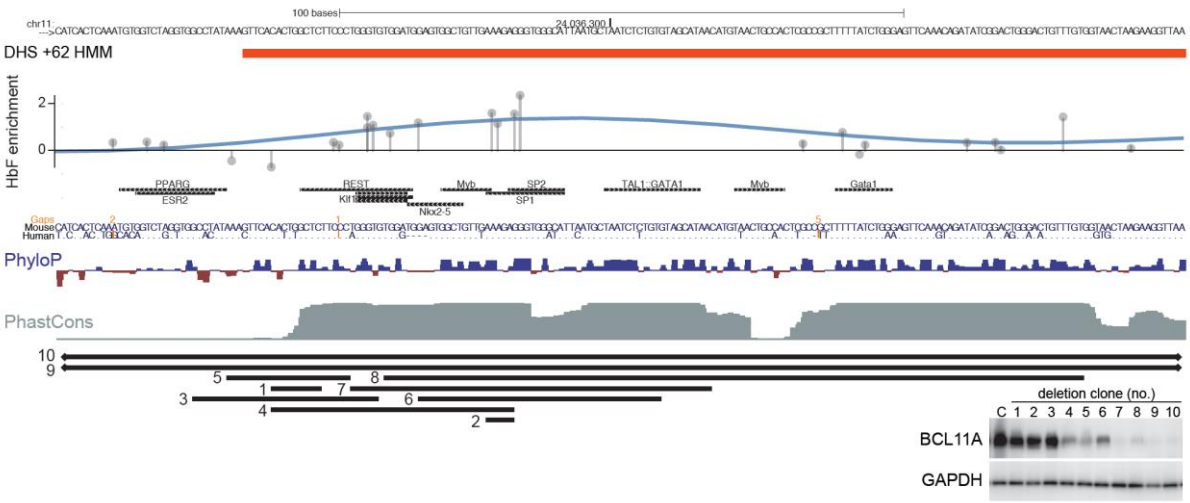
Supplemental Figure 3.6: Tiled pooled *in situ* CRISPR-Cas9 *Bcl11a* enhancer screen. a, Schematic of the mouse *Bcl11a* locus (mm9, transcription from left to right) with erythroid chromatin marks (*top*, dark blue H3K27ac from Kowalczyk *et al*[127], *middle*, light blue H3K27ac from Dogan *et al*[128], and *bottom*, black DNase I from Bauer *et al*[67]) and regions of primary sequence homology to the human DHSs displayed. Y-axes for H3K27ac tracks are both scaled to maximum 3.5 reads per million. Composite enhancer as previously defined[67]. **b**, Ranked enhancers in mouse erythroid precursors by H3K27ac signal intensity[127,128], with super-enhancers shaded. Super-enhancer associated genes indicated by Venn diagram. **c**, Strategy to knock-in by homology-directed repair the fluorescent protein mCherry into the mouse embryonic globin *Hbb-y* locus (encoding the εy embryonic globin chain). **d**, Distribution of NGG and NAG PAM sgRNAs mapped to genomic cleavage position with vertical lines representing cleavage sites for sgRNAs mapped to plus and minus strands. **e**, Distance to adjacent genomic cleavage position for NGG (*left*) and NAG (*right*) PAM sgRNAs. **f**, Representation of the 1,271 NGG and NAG sgRNAs within the plasmid pool by deep-sequencing. The median was 735 normalized reads and the 10th and 90th percentiles (indicated by the vertical dotted lines) ranged from 393 to 1,240 normalized reads. **g**, Library composition by target sequence and PAM restriction. **h**, mCherry expression upon exposure to Cas9 and an individual NGG sgRNA targeting *Bcl11a* exon 2 in MEL εγ:mCherry reporter cells. **i**, εγ:mCherry sort of library transduced cells. **j**, Control sgRNA enrichment. Boxes demonstrate 25th, median, and 75th percentiles and whiskers minimum and maximum values. **** $P < 0.0001$. **k**, Enrichment scores of NGG sgRNAs between four biological replicates.



Supplemental Figure 3.7: *Bcl11a* enhancer screen analyses. a, NGG sgRNA representation in plasmid pool and cells at conclusion of experiment (*left*), and in εy:mCherry-high and εy:mCherry-low pools (*right*), with dotted lines at x=y and x=8y. **b**, Quantile-quantile plots of NGG sgRNA εy enrichment scores. **c**, Cellular dropout scores of NGG sgRNAs relative to genomic cleavage position and repetitive elements. Nontargeting sgRNAs pseudo-mapped with 5 bp spacing. **d**, Correlation between cellular dropout and εy enrichment scores.



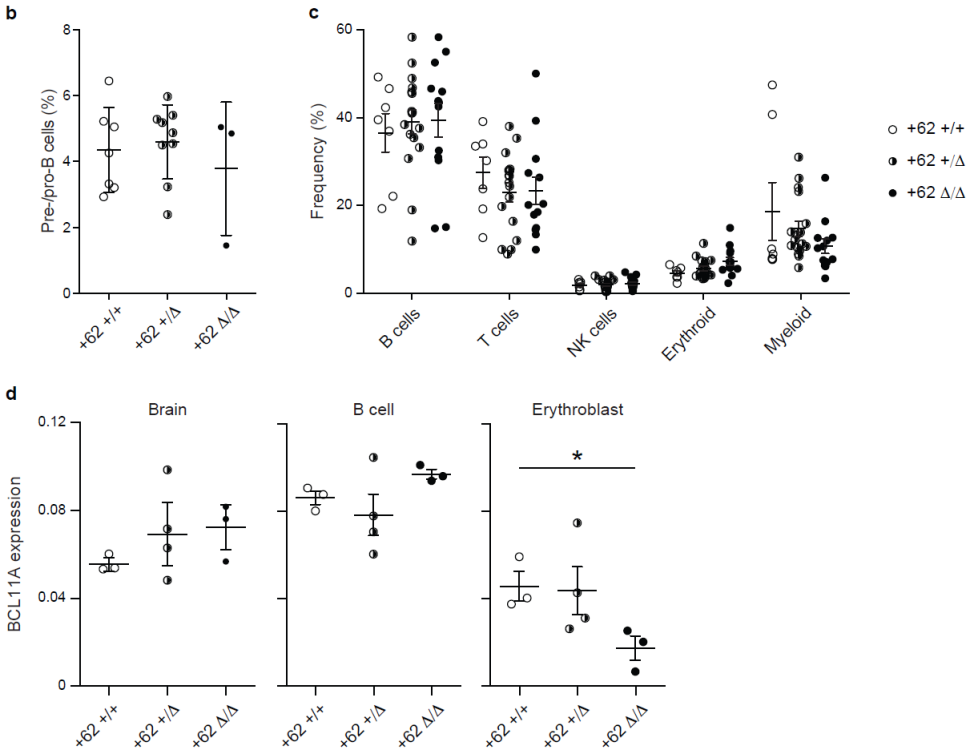
Supplemental Figure 3.8: Functional sequences at the *Bcl11a* erythroid enhancer. a-c, HMM segmentation of active functional states at m+55, m+58, and m+62 orthologs. HbF enrichment scores shown as gray lines and circles with blue line representing smoothed enrichment score. DNase I sequencing from mouse fetal liver erythroid precursors[67]. PhyloP (scale from -3.3 to 2.1) and PhastCons (from 0 to 1) estimates of evolutionary conservation among 30 vertebrates. **d**, *Top*, BCL11A expression determined by RT-qPCR displayed as a heatmap in 108 hemizygous m+62 ortholog deletion clones ordered by genomic position of deletion midpoint. Each bar demonstrates the genomic position of the deletion breakpoints and the associated color demonstrates the level of BCL11A expression. *Bottom*, BCL11A expression determined by RT-qPCR in 108 hemizygous m+62 ortholog deletion clones. Per nucleotide mean effect size was calculated as the mean fold change in BCL11A expression from all clones in which that nucleotide was deleted. Gray shading represents one s.d. The BCL11A expression data are shown with same x-axis as in Supplemental Figure 3.8c immediately above.



Supplemental Figure 3.9: Evaluation of the m+62 functional core. 200 bp at the functional core of the m+62 ortholog defined by HMM state. Enrichment scores shown as gray lines and circles with blue line representing smoothed enrichment score. JASPAR motifs ($P < 10^{-4}$) depicted with selected motifs annotated by TF name based on known erythroid-specific function or genomic position. Orthologous human sequences listed. PhyloP (scale from -3.3 to 2.1) and PhastCons (from 0 to 1) estimates of evolutionary conservation among 30 vertebrates. Individual numbered hemizygous deletion clones with indicated breakpoints were evaluated by BCL11A immunoblot (C, control). Clones 9 and 10 encompass the entire m+62 ortholog.

a

Genotype	Progeny (no.)	Percentage (expected)
+62 +/+	11	14.5 (25)
+62 +/Δ	45	59.2 (50)
+62 Δ/Δ	20	26.3 (25)
Total	76	100 (100)



Supplemental Figure 3.10: Requirement of *Bcl11a* erythroid enhancer during murine ontogeny. a, Progeny of heterozygous *Bcl11a* m+62 ortholog deletion intercrosses as compared to expected Mendelian ratio. **b,** Fraction of fetal liver comprised of B cell progenitors at E16.5 from various genotypes. **c**, Peripheral blood analysis from 4 week old mice to examine the frequency of various circulating hematopoietic lineages in *Bcl11a* m+62 ortholog deletion wild-type, heterozygous, and homozygous mice. **d**, BCL11A expression in β-YAC / +62 deletion mice (each symbol represents the mean expression from technical replicates from an individual mouse). * $P < 0.05$, error bars represent s.e.m.

Supplemental Table 3.1: sgRNA Sequences

sgRNA Target Gene or Region	Species	Sequence
Composite Enhancer 5' Target 1	Human	TGGAAAGGAGAACGGCCCGG
Composite Enhancer 5' Target 2	Human	TGAACACCCTCGTTAAAGGC
Composite Enhancer 5' Target 3	Human	AACACTAGCCCACATGCCAA
Composite Enhancer 3' Target 1	Human	GCCCACAGAGGCACGGTTAA
Composite Enhancer 3' Target 2	Human	AGGCACGGTTAATGGTGGCG
Composite Enhancer 3' Target 3	Human	CACAGGAAGCCATGGTCCTT
+55 5' Target 1	Human	GCACTGACGTAGGTAGTGAC
+55 5' Target 2	Human	ATAGGATATGGCACTGACGT
+55 3' Target 1	Human	CATTATCTTCTCTGGTCTCG
+55 3' Target 2	Human	ATACTGGGGAACACATTGTA
+58 5' Target 1	Human	TGAGCACATTCTTACGCCTA
+58 5' Target 2	Human	CTAGGCGTAAGAATGTGCTC
+58 3' Target 1	Human	GAACCCCCTATAAACTAGTC
+58 3' Target 2	Human	GGCAAACCAGACTAGTTTAT
+62 5' Target 1	Human	CAGGGGAGAACTCGGCATGA
+62 5' Target 2	Human	GATGGAGTTGGTTGACCGTA
+62 3' Target 1	Human	GGTAGGACCCAACACTACGC
+62 3' Target 2	Human	ATGCCTAGGGTGTTTTGACG
BCL11A Exon 2 Target 2	Human	TGAACCAGACCACGGCCCGT
BCL11A Exon 2 Target 3	Human	GCATCCAATCCCGTGGAGGT
+55 5' Target	Mouse	CACTGGCTTCCTGTTCTTGT
+55 3' Target	Mouse	AAGGTTTTCAAGGCAAATAA
+58 5' Target	Mouse	GTAATGGAGCCCGCATGCTG
+58 3' Target	Mouse	GCCAGTGTACAGGCAAGTAC
+62 5' Target	Mouse	TCGCTGCCTTCAGTTCTGCT
+62 3' Target	Mouse	TTATGGAACTCAGGAACTGC
Bcl11a Exon 2 Target	Mouse	GATGCCTTTTTCATCTCGAT
+62 Target 1	Mouse	ATTCCTTGAGTGTCATATAT
+62 Target 2	Mouse	TCTGGAATCACTATGTATAT

Supplemental Table 3.2: Oligonucleotides for deletion clone screening

Gene or Region	Species	Non-Deletion (ND) or Deletion (D)	CRISPR Pair	Orientation	Sequence
Composite Enhancer	Human	ND	5' Target 3	Forward	TGCTCCGAGCTTGTGAACTA
			3' Target 1	Reverse	TATCACAGGCTCCAGGAAGG
Composite Enhancer	Human	D	5' Target 3	Forward	TAGTTTGCTTCCCCCAATGA
			3' Target 1	Reverse	GCCAGGAAATTGGTGGTAGA
Composite Enhancer	Human	ND	5' Target 2	Forward	TGCTCCGAGCTTGTGAACTA
			3' Target 2	Reverse	TATCACAGGCTCCAGGAAGG
Composite Enhancer	Human	D	5' Target 2	Forward	GTGGGCAGTTACGTTTTCGT
			3' Target 2	Reverse	GCCAGGAAATTGGTGGTAGA
+55	Human	ND	5' Target 1 or 2	Forward	GGTCAGGGTGTTGCAGAGAT
			3' Target 1 or 2	Reverse	CACACCCTGTGATCTTGTGG
+55	Human	D	5' Target 1 or 2	Forward	GACTTAAACTGCCGCTCCTG
			3' Target 1 or 2	Reverse	GGGCCTCAGGCTCTTTATCT
+58	Human	ND	5' Target 1 or 2	Forward	CCCAGAGCTCAGTGAGATGA
			3' Target 1 or 2	Reverse	GGGAAAGGGCCTGATAACTT
+58	Human	D	5' Target 1 or 2	Forward	GAACAGAGACCACTACTGGCAAT
			3' Target 1 or 2	Reverse	CTCAGAAAAATGACAGCACCA
+62	Human	ND	5' Target 1 or 2	Forward	TTTGAAAGTACCAGCACAGCA
			3' Target 1 or 2	Reverse	CCCTCTGGCATCAAAATGAG
+62	Human	D	5' Target 1 or 2	Forward	AACAGACCCATGTGCTAGGC
			3' Target 1 or 2	Reverse	TGCTGAATTCCTGTAAAGTGAGG
+55	Mouse	ND	5' Target	Forward	GAGGTGACCAGGGTGTGAGT
			3' Target	Reverse	AAGAAGAGGCCCTGGACATT
+55	Mouse	D	5' Target	Forward	CATCTTAAGGCAAGAATCACT
			3' Target	Reverse	CCAGTCAATCCAAACCCTGT
+58	Mouse	ND	5' Target	Forward	TATTAATGCCCAGCCAGCTC
			3' Target	Reverse	GTGGTCCAGACCTAGCCAAG
+58	Mouse	D	5' Target	Forward	TTTGAGCAGGAGGGAATTTG
			3' Target	Reverse	ATAGGTGGTTGGGCTTCTCC
+62	Mouse	ND	5' Target	Forward	GGAGTGGCTGTTGAAAGAGG
			3' Target	Reverse	CACTCAAGGAATGCAAGCAA
+62	Mouse	D	5' Target	Forward	TACTTGGTGGCTTTCCCAAC
			3' Target	Reverse	AGATGGTCCTCTGCATCCAC

Supplemental Table 3.3: Oligonucleotides for inversion clone screening

Inverted Region	Species	CRISPR Pair	Orientation	Sequence
+55	Human	5’ Target 1 or 2	Forward	GACTTAAACTGCCGCTCCTG
		3’ Target 1 or 2	Forward	AGGCATCCAAAGGGAAGAAT
+55	Human	5’ Target 1 or 2	Reverse	ACTTCAGCCTCCAGCACTGT
		3’ Target 1 or 2	Reverse	CCACTGGAGTGGAACCAAGT
+58	Human	5’ Target 1 or 2	Forward	GGGATCAGAGGTGAACAGGA
		3’ Target 1 or 2	Forward	TGGACTTTGCACTGGAATCA
+58	Human	5’ Target 1 or 2	Reverse	TTGTTTACAGAGGGGCAACC
		3’ Target 1 or 2	Reverse	GGGGAAGGGGTATTGAATTG
+62	Mouse	5’ Target 1 or 2	Forward	AACAGACCCATGTGCTAGGC
		3’ Target 1 or 2	Forward	GAACCTGGGAGGCAGAAGAT
+62	Mouse	5’ Target 1 or 2	Reverse	TGTGTGGACTGCCTTTTCTG
		3’ Target 1 or 2	Reverse	TGTGGAGCTCTGGAATGATG

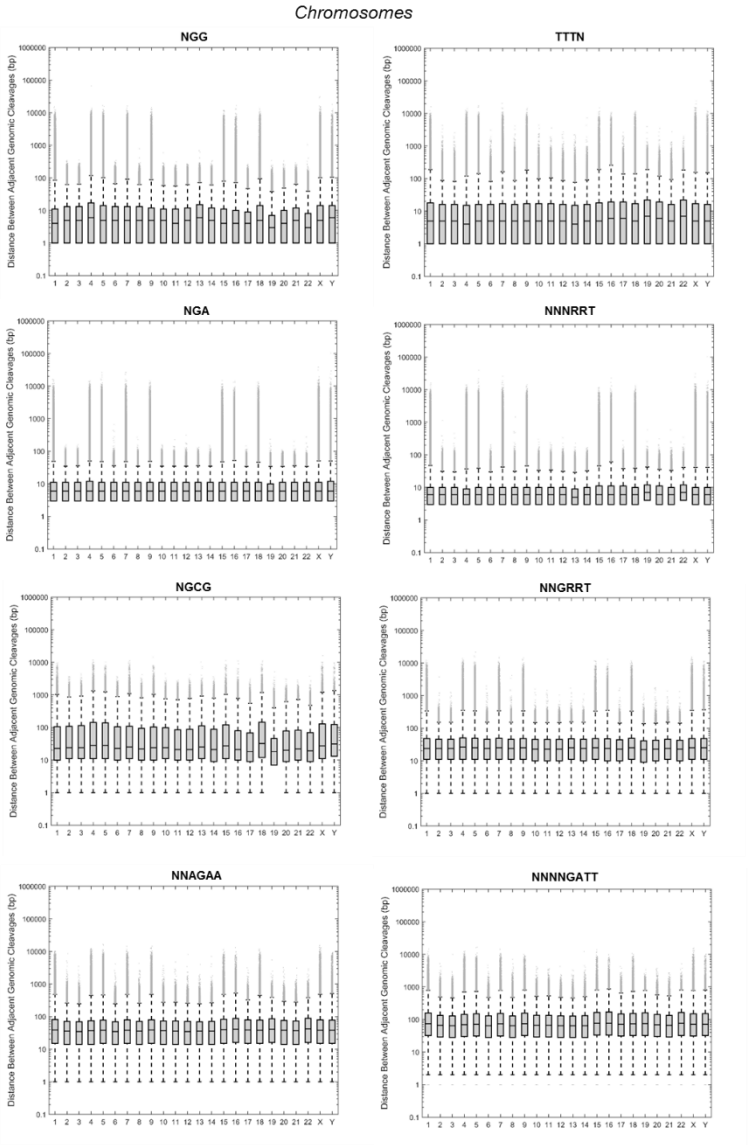
Supplemental Table 3.4: Oligonucleotides for mouse +62 deletion analysis

Region	Species	CRISPR Pair	Orientation	Sequence
+62	Mouse	Screen 0484	Forward	GGTAGTGTGGGGGTGGAGT
		Screen 0475	Reverse	TCAGCCTGTTCCCTCAGTG
+62	Mouse	Screen 0484	Forward	GGTAGTGTGGGGGTGGAGT
		Screen 2456	Reverse	TCAGCCTGTTCCCTCAGTG
+62	Mouse	Screen 0475	Forward	GGTAGTGTGGGGGTGGAGT
		Screen 0490	Reverse	TCAGCCTGTTCCCTCAGTG
+62	Mouse	Screen 0490	Forward	GGTAGTGTGGGGGTGGAGT
		+62 3' Target	Reverse	AGATGGTCCTCTGCATCCAC
+62	Mouse	Screen 0490	Forward	GGTAGTGTGGGGGTGGAGT
		+62 Target 1	Reverse	TCAGCCTGTTCCCTCAGTG
+62	Mouse	+62 5' Target	Forward	TACTTGGTGGCTTTCCCAAC
		Screen 0475	Reverse	TCAGCCTGTTCCCTCAGTG
+62	Mouse	+62 Target 2	Forward	ATGCTTGGTTGTCGCCTTAT
		Screen 0475	Reverse	CACTCAAGGAATGCAAGCAA

Supplemental Table 3.5: RT qPCR oligonucleotides

Gene	Species	Orientation	Sequence
GAPDH	Human	Forward	ACCCAGAAGACTGTGGATGG
		Reverse	TTCAGCTCAGGGATGACCTT
HBB	Human	Forward	CTGAGGAGAAGTCTGCCGTTA
		Reverse	AGCATCAGGAGTGGACAGAT
HBG	Human	Forward	TGGATGATCTCAAGGGCAC
		Reverse	TCAGTGGTATCTGGAGGACA
HBE	Human	Forward	GCAAGAAGGTGCTGACTTCC
		Reverse	ACCATCACGTTACCCAGGAG
HBD	Human	Forward	GAGGAGAAGACTGCTGTCAATG
		Reverse	AGGGTAGACCACCAGTAATCTG
BCL11A	Human	Forward	AACCCCAGCACTTAAGCAAA
		Reverse	GGAGGTCATGATCCCCTTCT
Gapdh	Mouse	Forward	TGGTGAAGGTCGGTGTGAAC
		Reverse	CCATGTAGTTGAGGTCAATGAAGG
β-Major	Mouse	Forward	TTTAACGATGGCCTGAATCACTT
		Reverse	CAGCACAATCACGATCATATTGC
Hbb-εy	Mouse	Forward	TGGCCTGTGGAGTAAGGTCAA
		Reverse	GAAGCAGAGGACAAGTTCCCA
Hbb-βh1	Mouse	Forward	TGGACAACCTCAAGGAGACC
		Reverse	ACCTCTGGGGTGAATTCCTT
Bcl11a	Mouse	Forward	AACCCCAGCACTTAAGCAAA
		Reverse	ACAGGTGAGAAGGTCGTGGT

CHAPTER 4 SUPPLEMENTAL MATERIAL

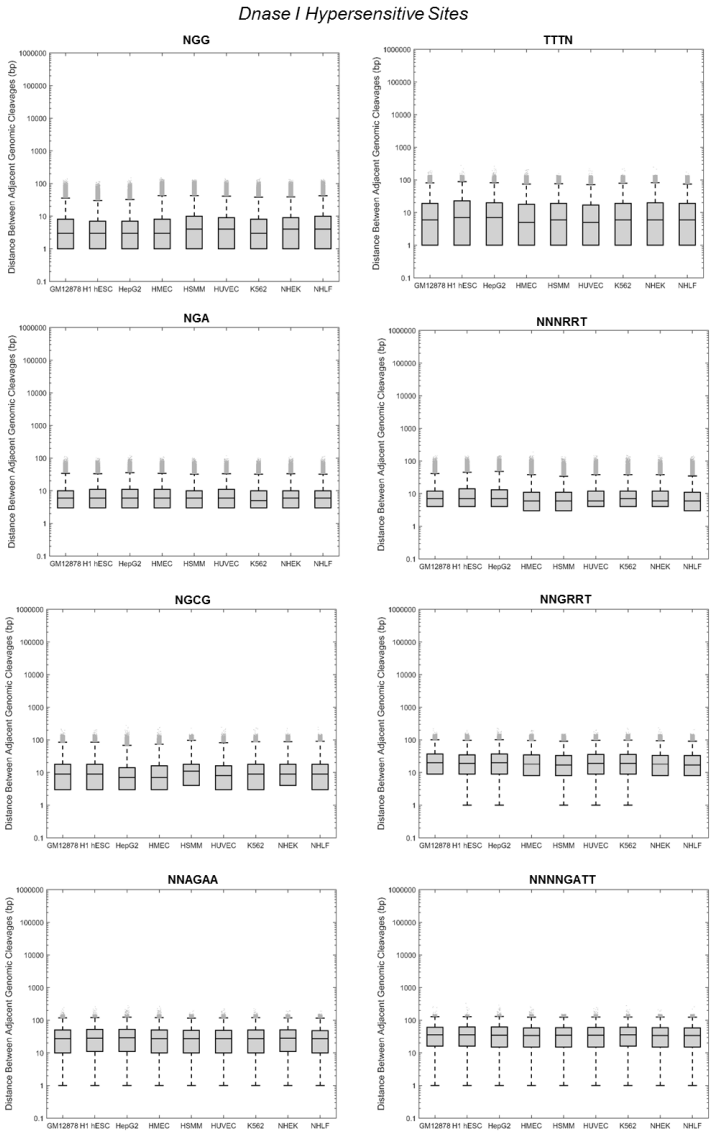


Supplemental Figure 4.1: Degree of saturation for 8 PAM sequences by chromosome. Distances between adjacent genomic cleavages to assess PAM availability and distribution by chromosome. For each box-and-whisker plot, the three lines of box represent the 25th, 50th and 75th percentile. The upper and lower whiskers represent the 99th and 1st percentile, respectively. Outliers, defined as above the 99th percentile or below the 1st percentile are plotted as individual points. Lower whiskers are omitted if the 1st percentile is 0.

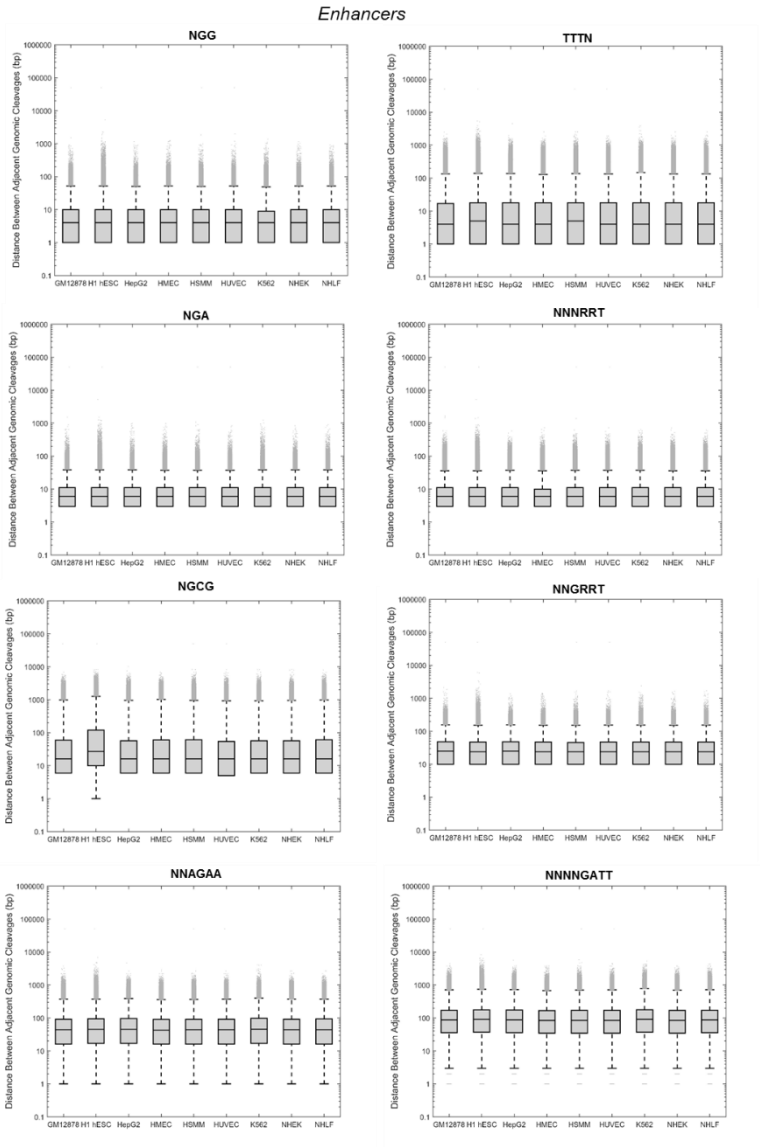
	Median (bp)																							
Chromosome	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**X**	**Y**
NGG	4	5	5	6	5	5	5	5	5	5	4	5	6	5	4	4	4	5	3	4	5	3	5	6
TTTN	5	5	5	4	5	5	5	5	5	5	5	5	4	5	5	6	6	5	7	6	5	7	5	5
NGA	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
NNNRRT	6	6	6	6	6	6	6	6	6	6	6	6	5	6	6	6	6	6	7	6	6	7	6	6
NGCG	23	24	24	28	28	23	25	22	24	24	21	21	25	21	27	21	18	32	18	20	22	19	27	31
NNGRRT	24	24	24	26	25	24	25	24	25	23	23	23	25	24	25	24	23	25	22	23	24	23	25	25
NNAGAA	39	36	35	37	38	35	38	36	39	37	36	35	36	36	40	41	39	38	41	38	37	42	38	38
NNNNGATT	74	66	65	70	72	65	73	65	74	67	67	65	64	65	77	77	71	73	75	69	66	76	72	72

	Mean (bp)																							
Chromosome	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**X**	**Y**
NGG	11.7	9.8	10.3	15.8	13.9	10.4	12.7	9.9	12.1	9.0	8.6	9.6	11.4	9.4	11.3	10.0	7.1	13.4	5.9	7.6	9.5	6.1	14.2	15.0
TTTN	17.5	12.2	11.8	14.5	15.5	11.7	16.2	12.2	17.0	13.1	13.5	12.2	11.2	12.5	17.8	20.2	15.9	15.6	19.9	14.9	12.3	19.7	17.0	16.9
NGA	11.1	8.3	8.3	11.6	11.3	8.3	11.2	8.3	11.2	8.2	8.1	8.2	8.5	8.3	10.9	10.9	8.0	11.0	7.9	8.0	8.4	8.0	12.0	12.6
NNNRRT	10.4	7.5	7.4	9.6	9.9	7.3	10.1	7.5	10.4	7.8	7.8	7.5	7.1	7.6	10.5	11.3	8.5	9.7	9.3	8.3	7.6	9.3	10.7	11.0
NGCG	108.5	97.6	104.5	143.5	134.5	99.1	113.4	89.5	107.3	88.8	80.0	85.1	104.4	86.2	115.8	81.0	63.5	136.0	45.4	72.4	79.0	61.2	131.4	136.0
NNGRRT	42.3	33.6	33.7	45.2	43.7	33.9	43.3	33.5	43.0	33.1	32.7	33.1	34.7	33.4	42.1	41.4	32.0	43.1	30.1	31.9	34.0	31.8	44.4	45.6
NNAGAA	69.4	52.9	51.6	65.4	66.7	51.7	68.4	52.4	69.5	54.5	54.4	52.1	51.5	53.4	70.6	75.1	59.7	66.7	66.7	57.1	54.5	67.7	69.7	71.1
NNNNGATT	129.4	98.3	95.5	118.4	122.6	96.6	125.9	96.5	128.7	101.5	101.8	96.7	94.2	98.8	133.0	137.5	114.1	123.7	126.8	107.0	100.1	129.1	125.3	129.4

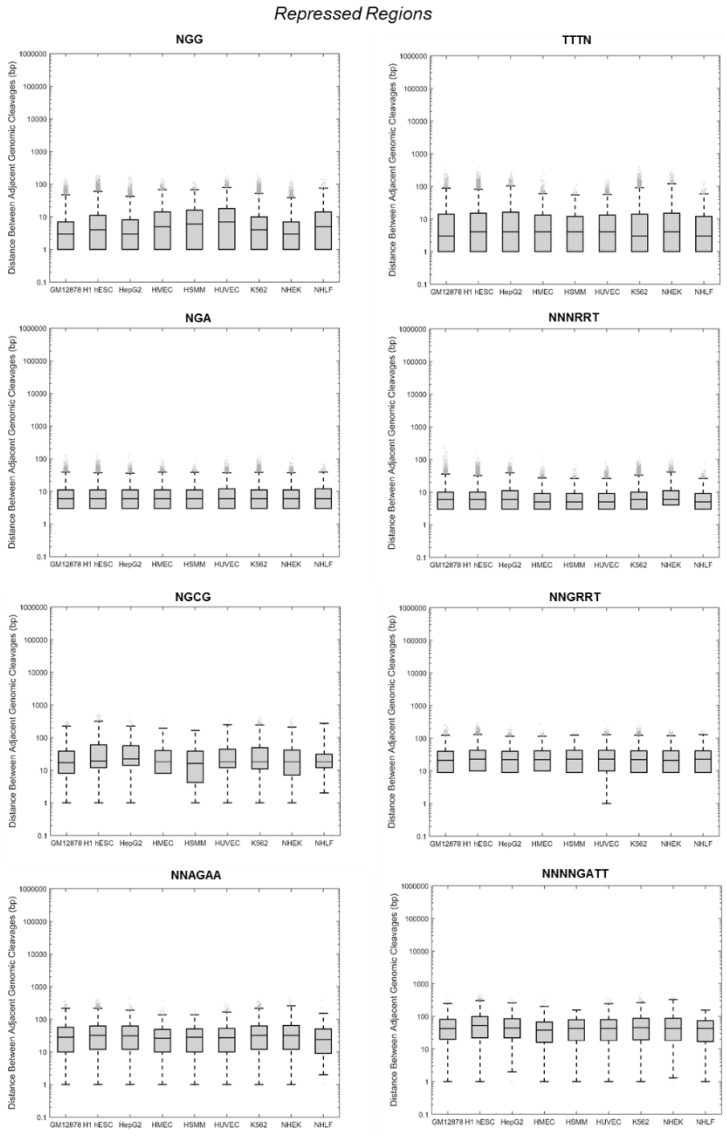
Supplemental Figure 4.2: Degree of saturation for 8 PAM sequences by chromosome.



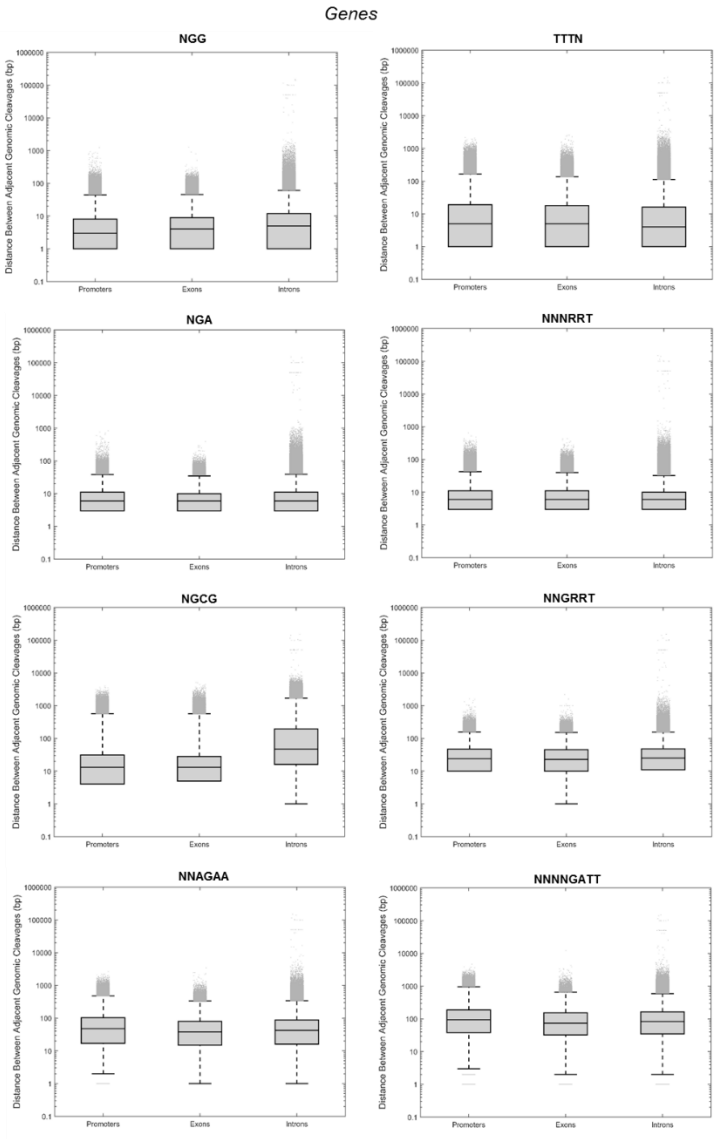
Supplemental Figure 4.3: Degree of saturation for 8 PAM sequences in DHS regions for 9 ENCODE cell lines. Distances between adjacent genomic cleavages to assess PAM availability and distribution within DHS regions. For each box-and-whisker plot, the three lines of box represent the 25th, 50th and 75th percentile. The upper and lower whiskers represent the 99th and 1st percentile, respectively. Outliers, defined as above the 99th percentile or below the 1st percentile are plotted as individual points. Lower whiskers are omitted if the 1st percentile is 0.



Supplementary Figure 4.4: Degree of saturation for 8 PAM sequences in enhancer regions for 9 ENCODE cell lines. Distances between adjacent genomic cleavages to assess PAM availability and distribution within enhancer regions. For each box-and-whisker plot, the three lines of box represent the 25th, 50th and 75th percentile. The upper and lower whiskers represent the 99th and 1st percentile, respectively. Outliers, defined as above the 99th percentile or below the 1st percentile are plotted as individual points. Lower whiskers are omitted if the 1st percentile is 0.



Supplemental Figure 4.5: Degree of saturation for 8 PAM sequences in repressed regions for 9 ENCODE cell lines. Distances between adjacent genomic cleavages to assess PAM availability and distribution within repressed regions. For each box-and-whisker plot, the three lines of box represent the 25th, 50th and 75th percentile. The upper and lower whiskers represent the 99th and 1st percentile, respectively. Outliers, defined as above the 99th percentile or below the 1st percentile are plotted as individual points. Lower whiskers are omitted if the 1st percentile is 0.



Supplemental Figure 4.6: Degree of saturation for 8 PAM sequences in RefSeq genes annotations. Distances between adjacent genomic cleavages to assess PAM availability and distribution within RefSeq gene annotated regions. For each box-and-whisker plot, the three lines of box represent the 25th, 50th and 75th percentile. The upper and lower whiskers represent the 99th and 1st percentile, respectively. Outliers, defined as above the 99th percentile or below the 1st percentile are plotted as individual points. Lower whiskers are omitted if the 1st percentile is 0.

a

Median distance between adjacent genomic cleavages in DHS (bp)									
Cell Type	**GM12878**	**H1 hESC**	**HepG2**	**HMEC**	**HSMM**	**HUVEC**	**K562**	**NHEK**	**NHLF**
NGG	3	3	3	3	4	4	3	4	4
TTTN	6	7	7	5	6	5	6	6	6
NGA	6	6	6	6	6	6	5	6	6
NNNRRT	7	7	7	6	6	6	7	6	6
NGCG	9	9	7	7	11	8	9	9	9
NNGRRT	20	19	20	18	17	19	19	18	17
NNAGAA	27	28	29	27	27	27	27	28	27
NNNNGATT	36	36	35	34	35	35	36	34	34

Mean distance between adjacent genomic cleavages in DHS (bp)									
Cell Type	**GM12878**	**H1 hESC**	**HepG2**	**HMEC**	**HSMM**	**HUVEC**	**K562**	**NHEK**	**NHLF**
NGG	6.0	5.2	5.1	6.7	7.6	6.8	6.4	6.6	7.4
TTTN	13.5	15.7	13.7	12.3	13.0	11.9	13.4	14.0	12.8
NGA	7.9	7.9	8.3	8.1	7.8	7.9	7.6	7.8	7.7
NNNRRT	9.3	10.2	10.2	8.5	8.4	8.8	9.0	8.8	8.4
NGCG	13.8	14.5	10.9	11.6	17.1	13.1	14.4	14.8	14.6
NNGRRT	26.0	24.8	26.2	24.7	23.7	25.3	25.6	23.9	23.5
NNAGAA	33.6	35.2	36.2	33.7	33.5	33.6	33.7	34.3	33.1
NNNNGATT	41.9	42.1	41.9	39.8	40.3	40.8	41.6	40.5	40.1

b

Median distance between adjacent genomic cleavages in enhancers (bp)									
Cell Type	**GM12878**	**H1 hESC**	**HepG2**	**HMEC**	**HSMM**	**HUVEC**	**K562**	**NHEK**	**NHLF**
NGG	4	4	4	4	4	4	4	4	4
TTTN	4	5	4	4	5	4	4	4	4
NGA	6	6	6	6	6	6	6	6	6
NNNRRT	6	6	6	6	6	6	6	6	6
NGCG	16	27	16	16	16	16	16	16	16
NNGRRT	25	24	25	24	24	24	24	24	24
NNAGAA	44	45	45	43	44	44	46	44	44
NNNNGATT	87	90	88	84	85	85	91	85	87

Mean distance between adjacent genomic cleavages in enhancer (bp)									
Cell Type	**GM12878**	**H1 hESC**	**HepG2**	**HMEC**	**HSMM**	**HUVEC**	**K562**	**NHEK**	**NHLF**
NGG	7.8	7.9	7.7	8.1	7.7	7.8	7.5	7.8	7.8
TTTN	14.7	15.4	15.0	14.6	15.3	14.8	15.6	14.9	14.9
NGA	8.3	8.3	8.4	8.3	8.2	8.3	8.3	8.3	8.3
NNNRRT	8.3	8.2	8.3	8.2	8.3	8.3	8.4	8.2	8.3
NGCG	81.8	126.3	78.9	84.2	81.5	76.0	76.8	79.0	82.2
NNGRRT	34.3	34.1	34.4	33.9	33.4	34.0	33.9	33.9	34.1
NNAGAA	69.4	70.6	71.9	67.7	68.5	69.3	72.9	69.4	69.6
NNNNGATT	131.9	136.8	133.2	127.6	130.1	130.3	138.7	129.3	131.9

c

Median distance between adjacent genomic cleavages in repressed regions (bp)									
Cell Type	**GM12878**	**H1 hESC**	**HepG2**	**HMEC**	**HSMM**	**HUVEC**	**K562**	**NHEK**	**NHLF**
NGG	3	4	3	5	6	7	4	3	5
TTTN	3	4	4	4	4	4	3	4	3
NGA	6	6	6	6	6	6	6	6	6
NNNRRT	6	6	6	5	5	5	6	6	5
NGCG	17	19	22	18	16	18	18	18	18
NNGRRT	21	23	22	22	23	23	22	21	22.5
NNAGAA	28	32	31	26	28	27	32	32	24
NNNNGATT	42	53	43.5	38	43	43	45	43	43

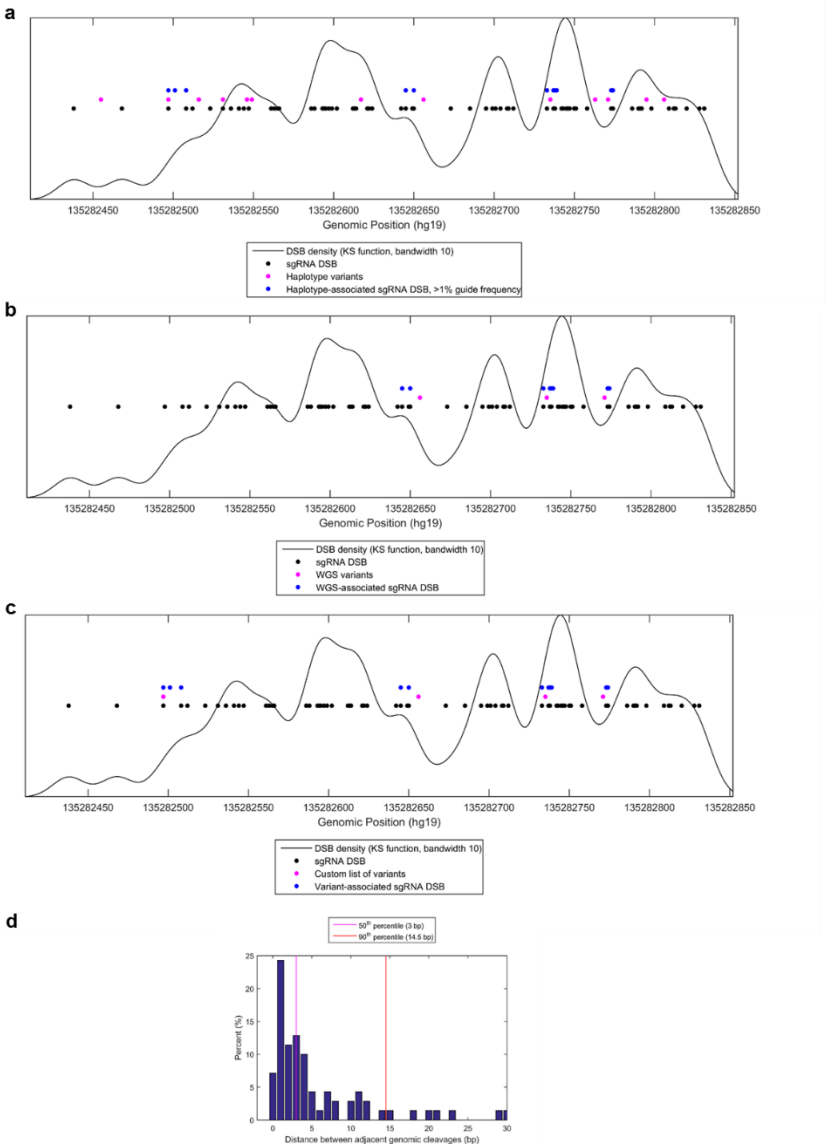
Mean distance between adjacent genomic cleavages in repressed regions (bp)									
Cell Type	**GM12878**	**H1 hESC**	**HepG2**	**HMEC**	**HSMM**	**HUVEC**	**K562**	**NHEK**	**NHLF**
NGG	6.2	8.9	6.5	10.8	11.6	13.0	7.9	5.8	11.2
TTTN	10.8	11.1	12.5	9.1	8.8	8.8	11.3	12.7	8.6
NGA	8.3	8.3	8.0	8.6	8.6	8.7	8.3	8.3	8.9
NNNRRT	8.1	7.6	8.8	6.9	6.8	6.8	7.9	9.0	6.9
NGCG	32.4	48.4	43.5	31.3	28.1	37.5	39.5	32.3	30.0
NNGRRT	28.8	30.6	28.5	29.3	29.9	31.0	29.2	28.9	29.4
NNAGAA	41.5	45.1	43.8	34.8	35.8	37.5	45.2	48.1	35.4
NNNNGATT	58.2	70.9	60.1	49.3	51.6	56.7	62.8	62.9	51.2

d

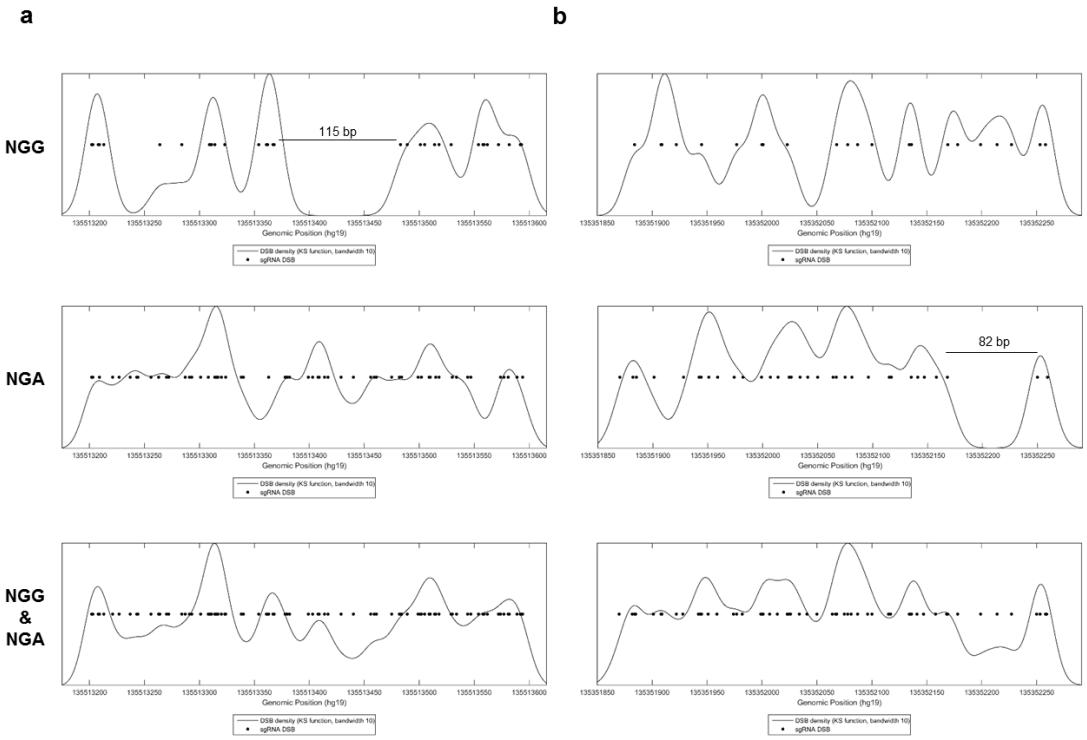
Median distance between adjacent genomic cleavages in genes (bp)			
Genes	**Promoters**	**Exons**	**Introns**
NGG	3	4	5
TTTN	5	5	4
NGA	6	6	6
NNNRRT	6	6	6
NGCG	13	13	47
NNGRRT	24	23	25
NNAGAA	48	38	42
NNNNGATT	95	74	83

Mean distance between adjacent genomic cleavages in genes (bp)			
Genes	**Promoters**	**Exons**	**Introns**
NGG	6.5	6.9	9.3
TTTN	17.7	15.5	13.0
NGA	8.2	7.9	8.5
NNNRRT	9.0	8.9	7.7
NGCG	45.0	41.7	185.2
NNGRRT	34.3	33.1	34.8
NNAGAA	79.3	60.1	65.2
NNNNGATT	151.5	117.3	121.3

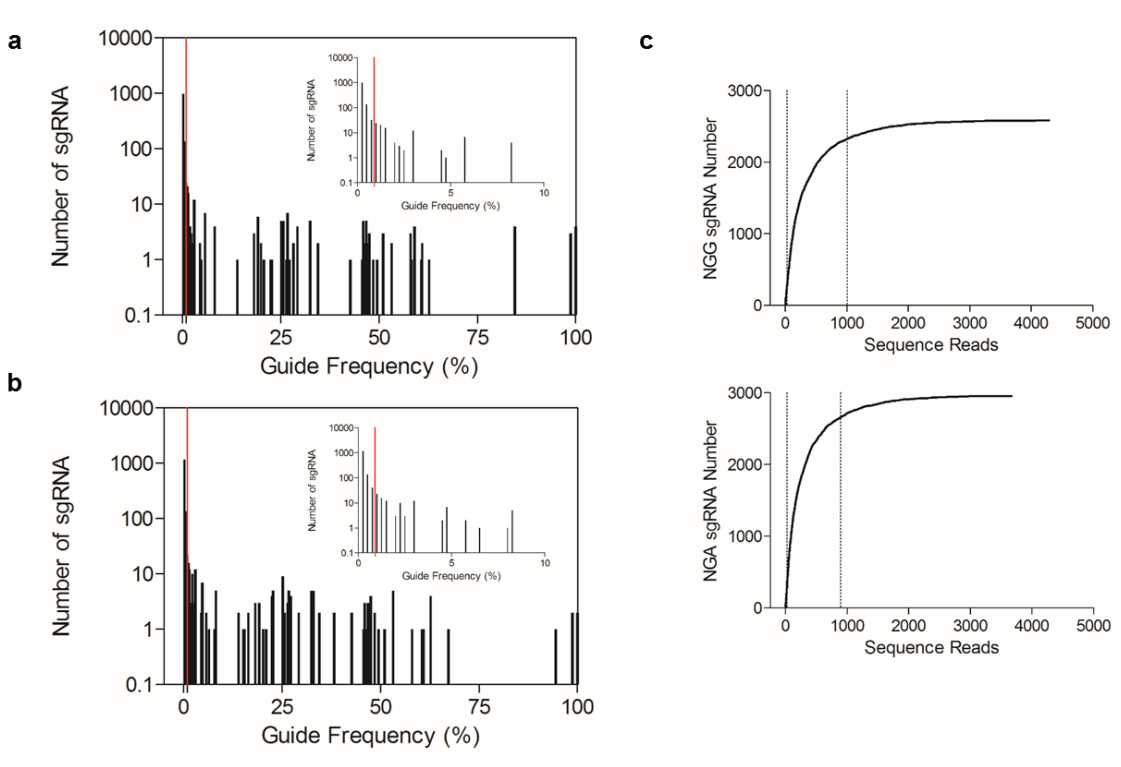
Supplemental Figure 4.7: Degree of saturation for 8 PAM sequences in (a) DHS, (b) enhancer, and (c) repressed regions for 9 ENCODE cell lines as well as (d) RefSeq gene annotations.



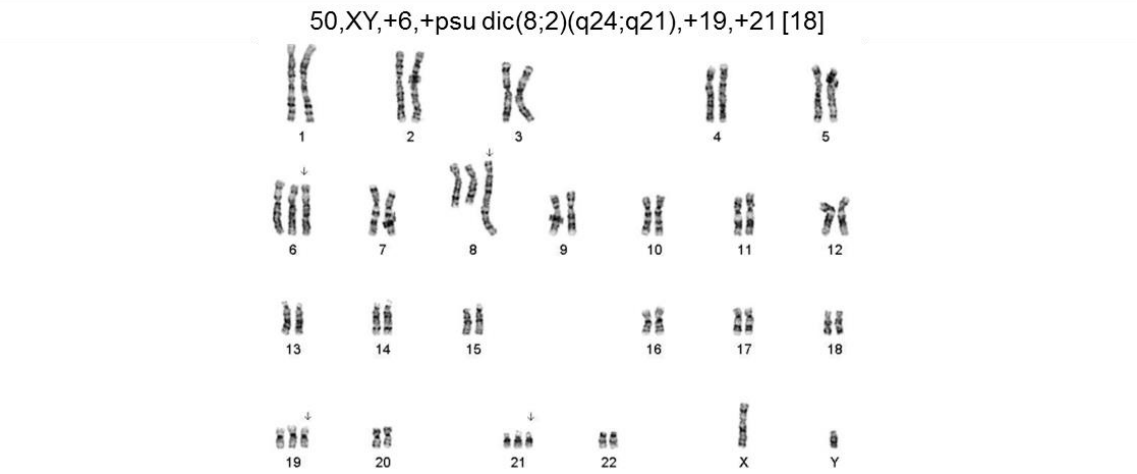
Supplemental Figure 4.8: Representative output figures from DNA Striker analysis using haplotype data, whole genome sequencing (WGS), and a custom list of variants for chr6:135282411-135282852 (hg19). a, Representative plot displaying location of non-variant sgRNA (*black*), haplotype variants (*pink*), and haplotype-associated sgRNA with a guide frequency ≥1% frequency (*blue*). The double strand break density is estimated using the KS density function (*black line*). **b**, Representative plot displaying location of non-variant sgRNA (*black*), WGS variants (*pink*), and WGS-associated sgRNA (*blue*). The double strand break density is estimated using the KS density function (*black line*). **c**, Representative plot displaying location of non-variant sgRNA (*black*), custom list of common variants (*pink*), and variant-associated sgRNA (*blue*). The double strand break density is estimated using the KS density function (*black line*). The custom list of variants represents all SNPs in the region with a minor allele frequency of ≥1%. **d**, Representative plot quantifying degree of saturation through determining gaps between adjacent genomic cleavages. The 50th percentile (*pink*) and 90th percentile (*red*) are indicated by vertical lines.



Supplemental Figure 4.9: Combining multiple nucleases reduces gap distance between adjacent cleavages. a, NGG-restricted sgRNA library includes a gap between adjacent genomic cleavages of 115 bp in HBS1L-MYB DHS (chr6:135513195-135513596, hg19) with 50th and 90th percentile gap distances of 5 bp and 27 bp, respectively. The combination of NGG- and NGA-restricted sgRNA reduces the maximum gap to 17 bp with 50th and 90th percentile gap distances of 3 bp and 11 bp, respectively. **b**, NGA-restricted sgRNA library includes a gap between adjacent genomic cleavages of 82 bp in HBS1L-MYB DHS (chr6:135351870-135352271, hg19) with 50th and 90th percentile gap distances of 8 bp and 18 bp, respectively. The combination of NGG- and NGA-restricted sgRNA reduces maximum gap to 23 bp with 50th and 90th percentile gap distances of 5 bp and 15 bp, respectively.



Supplemental Figure 4.10: NGG and NGA pooled saturating mutagenesis screen of the HBS1L-MYB intergenic region. a, Altered sgRNA or PAM-creation sgRNA in the HBS1L-MYB intergenic region as a result of variants from the 1,000 Genomes Project sorted by frequency. The cutoff for inclusion in the library was a frequency of 1%. **b**, Representation of both NGG (n = 2,585) and NGA (n = 2,957) sgRNA within the plasmid pool by deep sequencing. The median was 187 normalized reads and the 10th and 90th percentiles (indicated by the vertical dotted lines) ranged from 25 to 1,009 normalized reads for NGG sgRNA. The median was 162 normalized reads and the 10th and 90th percentiles (indicated by the vertical dotted lines) ranged from 21 to 896 normalized reads for NGA sgRNA.



Supplemental Figure 4.11: HUDEP-2 karyotype.

REFERENCES

1. Bauer, D. E. & Orkin, S. H. Hemoglobin switching's surprise: the versatile transcription factor BCL11A is a master repressor of fetal hemoglobin. *Curr. Opin. Genet. Dev.* **33,** 62–70 (2015).
2. Ingram, V. A Specific Chemical Difference Between the Globins of Normal Human and Sickle-Cell Anæmia Hæmoglobin. *Nature* **178,** 792–794 (1956).
3. Orkin, S. H. & Higgs, D. R. Sickle cell disease at 100 years. *Science* **329,** 291–2 (2010).
4. Orkin, S. H. *et al.* Linkage of β-thalassaemia mutations and β-globin gene polymorphisms with DNA polymorphisms in human β-globin gene cluster. *Nature* **296,** 627–631 (1982).
5. Orkin, S. H. & Kazazian, H. H. The mutation and polymorphism of the human β-globin gene and its surrounding DNA. *Ann. Rev. Genet.* **18,** 131–171 (1984).
6. Wong, C. *et al.* On the origin and spread of β-thalassemia: recurrent observation of four mutations in different ethnic groups. *Proc. Natl. Acad. Sci. U. S. A.* **83,** 6529–32 (1986).
7. Kazazian, H. H. *et al.* Molecular characterization of seven β-thalassemia mutations in Asian Indians. *EMBO J.* **3,** 593–596 (1984).
8. Antonarakis, S. E., Kazazian, H. H. & Orkin, S. H. DNA polymorphism and molecular pathology of the human globin gene clusters. *Hum. Genet.* 1–14 (1985). doi:10.1007/BF00295521
9. Lie-Injo, L. *et al.* β-thalassemia mutations in indonesia and their linkage to β haplotypes. *Am. J. Hum. Genet.* **45,** 971–975 (1989).
10. Platt, O. S. *et al.* Hydroxyurea enhances fetal hemoglobin production in sickle cell anemia. *J. Clin. Invest.* **74,** 652–656 (1984).
11. Letvin, N., Linch, D., Beardsley, G., McIntyre, K. & Nathan, D. Augmentation of Fetal-Hemoglobin Production in Anemic Monkeys by Hydroxyurea. *N. Engl. J. Med.* **310,** 869–73 (1984).
12. Charache, S. *et al.* Effect of Hydroxyurea on the Frequency of Painful Crises in Sickle

Cell. *N. Engl. J. Med.* **332,** 1317–1322 (1995).

13. Platt, O. S. Hydroxyurea for the treatment of sickle cell anemia. *N. Engl. J. Med.* **358,** 1362–9 (2008).
14. Bunn, F. Pathogenesis and Treatment of Sickle Cell Disease. *N. Engl. J. Med.* **337,** 762–9 (1997).
15. Rees, D. C., Williams, T. N. & Gladwin, M. T. Sickle-cell disease. *Lancet* **376,** 2018–31 (2010).
16. Rund, D. & Rachmilewitz, E. β-thalassemia. *N. Engl. J. Med.* **353,** 1135–46 (2005).
17. Platt, O. *et al.* Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N. Engl. J. Med.* **330,** 1639–1644 (1994).
18. Platt, O. *et al.* Pain in sickle cell disease: rates and risk factors. *N. Engl. J. Med.* **325,** 11–16 (1991).
19. Castro, O. *et al.* The acute chest syndrome in sickle cell disease: incidence and risk factors. The Cooperative Study of Sickle Cell Disease. *Blood* **84,** 643–649 (1994).
20. Musallam, K. M., Taher, A. T., Cappellini, M. D. & Sankaran, V. G. Clinical experience with fetal hemoglobin induction therapy in patients with β-thalassemia. *Blood* **121,** 2199–2212 (2013).
21. Galanello, R. *et al.* Amelioration of Sardinian βo thalasaemia by genetic modifiers. *Blood* **114,** 3935–3937 (2009).
22. Watson, J. A Study of Sickling of Young Erythrocytes in Sickle Cell Anemia. *Blood* **3,** 465–469 (1948).
23. Herman EC Jr, Conley, C. Hereditary Persistence of Fetal Hemoglobin. A Family Study. *Am. J. Med.* **29,** 9–17 (1960).
24. Cavazzana-Calvo, M. *et al.* Transfusion independence and HMGA2 activation after gene therapy of human β-thalassaemia. *Nature* **467,** 318–22 (2010).
25. Hoban, M. D., Orkin, S. H. & Bauer, D. E. Genetic Treatment of a Molecular Disorder:

Gene Therapy Approaches to Sickle Cell Disease. *Blood* (2016).

26. Kim, Y. G., Cha, J. & Chandrasegaran, S. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 1156–1160 (1996).
27. Smith, J., Berg, J. M. & Chandrasegaran, S. A detailed study of the substrate specificity of a chimeric restriction enzyme. *Nucleic Acids Res.* **27,** 674–681 (1999).
28. Bibikova, M., Golic, M., Golic, K. G. & Carroll, D. Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. *Genetics* **161,** 1169–1175 (2002).
29. Bibikova, M., Beumer, K., Trautman, J. K. & Carroll, D. Enhancing gene targeting with designed zinc finger nucleases. *Science* **300,** 764 (2003).
30. Porteus, M. H. & Baltimore, D. Chimeric nucleases stimulate gene targeting in human cells. *Science* **300,** 763 (2003).
31. Urnov, F. D. *et al.* Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435,** 646–651 (2005).
32. Moehle, E. A. *et al.* Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 3055–3060 (2007).
33. Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.* **11,** 636–46 (2010).
34. Miller, J. C. *et al.* An improved zinc-finger nuclease architecture for highly specific genome editing. *Nat. Biotechnol.* **25,** 778–85 (2007).
35. Li, H. *et al.* In vivo genome editing restores haemostasis in a mouse model of haemophilia. *Nature* **475,** 217–221 (2011).
36. Boch, J. *et al.* Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326,** 1509–1512 (2009).
37. Moscou, M. & Bogdanove, A. A simple cipher governs DNA recognition by TAL effectors. *Science* **326,** 1501 (2009).

38. Christian, M. *et al.* Targeting DNA Double-Strand Breaks with TAL Effector Nucleases. *Genetics* **186,** 757–761 (2010).

39. Li, T. *et al.* TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Res.* **39,** 359–372 (2011).

40. Miller, J. C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.* **29,** 143–148 (2011).

41. Stoddard, B. L. Homing Endonucleases: From Microbial Genetic Invaders to Reagents for Targeted DNA Modification. *Structure* **19,** 7–15 (2011).

42. Smith, J. *et al.* A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. *Nucleic Acids Res.* **34,** e149 (2006).

43. Silva, G. *et al.* Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy. *Curr. Gene Ther.* **11,** 11–27 (2011).

44. Thierry, A. & Dujon, B. Nested chromosomal fragmentation in yeast using the meganuclease I-Sce I: a new method for physical mapping of eukaryotic genomes. *Nucleic Acids Res.* **20,** 5625–5631 (1992).

45. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315,** 1709–1712 (2007).

46. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–21 (2012).

47. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* 819–23 (2013).

48. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–6 (2013).

49. Hsu, P. D., Lander, E. S. & Zhang, F. Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell* **157,** 1262–1278 (2014).

50. Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas

System. *Cell* **163,** 759–71 (2015).

51. Tsai, S. Q. *et al.* Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.* **32,** 569–76 (2014).

52. Wyvekens, N., Topkar, V. V., Khayter, C., Joung, J. K. & Tsai, S. Q. Dimeric CRISPR RNA-Guided FokI-dCas9 Nucleases Directed by Truncated gRNAs for Highly Specific Genome Editing. *Hum. Gene Ther.* **26,** 425–431 (2015).

53. Ran, F. A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520,** 186–91 (2015).

54. Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods* **10,** 1116–21 (2013).

55. Shi, J. *et al.* Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat. Biotechnol.* **33,** 661–7 (2015).

56. Ding, Q. *et al.* Enhanced efficiency of human pluripotent stem cell genome editing through replacing TALENs with CRISPRs. *Cell Stem Cell* **12,** 393–4 (2013).

57. Tebas, P. *et al.* Gene Editing of CCR5 in Autologous CD4 T Cells of Persons Infected with HIV. *N. Engl. J. Med.* **370,** 901–910 (2014).

58. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* **32,** 347–55 (2014).

59. Bibikova, M. *et al.* Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. *Mol. Cell. Biol.* **21,** 289–297 (2001).

60. Canver, M. C. *et al.* Characterization of Genomic Deletion Efficiency Mediated by Clusted Regularly Interspaced Palindromic Repeats (CRISPR)/Cas9 Nuclease System in Mammalian Cells. *J. Biol. Chem.* **289,** 21312–21324 (2014).

61. Maddalo, D. *et al.* In vivo engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system. *Nature* **516,** 423–7 (2014).

62. Choi, P. S. & Meyerson, M. Targeted genomic rearrangements using CRISPR/Cas

technology. *Nat. Commun.* **5,** 3728 (2014).

63. Blasco, R. B. *et al.* Simple and Rapid In Vivo Generation of Chromosomal Rearrangements using CRISPR/Cas9 Technology. *Cell Rep.* **9,** 1219–1227 (2014).
64. Xiao, A. *et al.* Chromosomal deletions and inversions mediated by TALENs and CRISPR/Cas in zebrafish. *Nucleic Acids Res.* **41,** e141 (2013).
65. Gupta, A. *et al.* Targeted chromosomal deletions and inversions in zebrafish. *Genome Res.* **23,** 1008–17 (2013).
66. Lee, H. J., Kim, E. & Kim, J. S. Targeted chromosomal deletions in human cells using zinc finger nucleases. *Genome Res.* **20,** 81–9 (2010).
67. Bauer, D. E. *et al.* An Erythroid Enhancer of BCL11A Subject to Genetic Variation Determines Fetal Hemoglobin Level. *Science* **342,** 253–257 (2013).
68. Ran, F. A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154,** 1380–9 (2013).
69. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8,** 2281–308 (2013).
70. Andrey, G. *et al.* Deletions, Inversions, Duplications: Engineering of Structural Variants using CRISPR/Cas in Mice. *Cell Rep.* **10,** 833–839 (2015).
71. Park, C.-Y. *et al.* Functional Correction of Large Factor VIII Gene Chromosomal Inversions in Hemophilia A Patient-Derived iPSCs Using CRISPR-Cas9. *Cell Stem Cell* **17,** 213–20 (2015).
72. Li, J. *et al.* Efficient inversions and duplications of mammalian regulatory DNA elements and gene clusters by CRISPR/Cas9. *J. Mol. Cell Biol.* **7,** 284–98 (2015).
73. Zhang, L. *et al.* Large Genomic Fragment Deletions and Insertions in Mouse Using CRISPR/Cas9. *PLoS One* **10,** e0120396 (2015).
74. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–23 (2013).

75. Sampson, T. R., Saroj, S. D., Llewellyn, A. C., Tzeng, Y. L. & Weiss, D. S. A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature* **497,** 254–7 (2013).

76. Xie, K. & Yang, Y. RNA-guided genome editing in plants using a CRISPR-Cas system. *Mol. Plant* **6,** 1975–83 (2013).

77. Bassett, A. R. & Liu, J. L. CRISPR/Cas9 and Genome Editing in Drosophila. *J. Genet. Genomics* **41,** 7–19 (2014).

78. Waaijers, S. *et al.* CRISPR/Cas9-targeted mutagenesis in Caenorhabditis elegans. *Genetics* **195,** 1187–91 (2013).

79. Yang, H. *et al.* One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. *Cell* **154,** 1370–9 (2013).

80. Guo, X. *et al.* Efficient RNA/Cas9-mediated genome editing in Xenopus tropicalis. *Development* 707–714 (2014).

81. Niu, Y. *et al.* Generation of Gene-Modified Cynomolgus Monkey via Cas9/RNA-Mediated Gene Targeting in One-Cell Embryos. *Cell* **156,** 836–843 (2014).

82. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* **8,** 2180–96 (2013).

83. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343,** 80–4 (2014).

84. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343,** 84–7 (2014).

85. Zhou, Y. *et al.* High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* (2014).

86. Schwank, G. *et al.* Functional Repair of CFTR by CRISPR/Cas9 in Intestinal Stem Cell Organoids of Cystic Fibrosis Patients. *Cell Stem Cell* **13,** 653–8 (2013).

87. Wu, Y. *et al.* Correction of a Genetic Disease in Mouse via Use of CRISPR-Cas9. *Cell*

Stem Cell **13,** 659–62 (2013).

88. Zhou, J. *et al.* Dual sgRNAs facilitate CRISPR/Cas9-mediated mouse genome targeting. *FEBS J.* **281,** 1717–1725 (2014).

89. Mamaeva, S. E. & Tsvileneva, N. N. A study of chromosome content of Friend virus-induced mouse erythroleukemia cells (clone M2) via karyotype reconstruction. *Cancer Genet. Cytogenet.* **16,** 199–205 (1985).

90. Miller, D. A., Tantravahi, R., Newman, B., Dev, V. G. & Miller, O. J. Karyotype of Friend virus-induced mouse erythroleukemia cells. *Cancer Genet. Cytogenet.* **1,** 103–113 (1979).

91. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–11 (2013).

92. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31,** 822–826 (2013).

93. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31,** 827–32 (2013).

94. Wu, X. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* (2014).

95. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32,** 279–284 (2014).

96. Guilinger, J. P., Thompson, D. B. & Liu, D. R. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat. Biotechnol.* (2014).

97. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–5 (2012).

98. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27,** 299–308 (1981).

99. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*

457, 854–858 (2009).

100. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **488,** 75–82 (2012).

101. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–61 (2014).

102. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).

103. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 21931–21936 (2010).

104. Xu, J. *et al.* Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev. Cell* **23,** 796–811 (2012).

105. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).

106. Parker, S. C. J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 17921–6 (2013).

107. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153,** 307–319 (2013).

108. Paul, D. S. *et al.* Maps of open chromatin guide the functional follow-up of genome-wide association signals: Application to hematological traits. *PLoS Genet.* **7,** (2011).

109. Hardison, R. C. Variable evolutionary signatures at the heart of enhancers. *Nat. Genet.* **42,** 734–735 (2010).

110. Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346,** 1007–1012 (2014).

111. Villar, D. *et al.* Enhancer Evolution across 20 Mammalian Species. *Cell* **160,** 554–566 (2015).

112. Pennacchio, L. a *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444,** 499–502 (2006).

113. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30,** 271–277 (2012).

114. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30,** 265–270 (2012).

115. Sexton, T. & Cavalli, G. The Role of Chromosome Domains in Shaping the Functional Genome. *Cell* **160,** 1049–1059 (2015).

116. Bender, M., Bulger, M., Close, J. & Groudine, M. Beta-globin gene switching and DNase I sensitivity of the endogenous beta-globin locus in mice do not require the locus control region. *Mol. Cell* **5,** 387–393 (2000).

117. Johnson, K. D. *et al.* Cis-element mutated in GATA2-dependent immunodeficiency governs hematopoiesis and vascular integrity. *J. Clin. Invest.* **122,** 3692–3704 (2012).

118. Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.* (2013).

119. Zhou, Y. *et al.* High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* **509,** 487–91 (2014).

120. Gröschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in Leukemia. *Cell* **157,** 369–381 (2014).

121. Mansour, M. R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* 10–15 (2014).

122. Kurita, R. *et al.* Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells. *PLoS One* **8,** e59890 (2013).

123. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for

CRISPR screening. *Nat. Methods* **11,** 783–784 (2014).

124. Chen, S. *et al.* Genome-wide CRISPR Screen in a Mouse Model of Tumor Growth and Metastasis. *Cell* **160,** 1246–60 (2015).

125. Giarratana, M. *et al.* Proof of principle for transfusion of in vitro generated red blood cells. *Blood* **118,** 5071–5079 (2011).

126. Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42,** e168 (2014).

127. Kowalczyk, M. S. *et al.* Intragenic Enhancers Act as Alternative Promoters. *Mol. Cell* **45,** 447–458 (2012).

128. Dogan, N. *et al.* Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin* **8,** 1–21 (2015).

129. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–1018 (2011).

130. Mathelier, A. *et al.* JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42,** 142–147 (2014).

131. Porcu, B. S. *et al.* The human β globin locus introduced by YAC transfer exhibits a specific and reproducible pattern of developmental regulation in transgenic mice. *Blood* **90,** 4602–4609 (1997).

132. Weber, K., Bartsch, U., Stocking, C. & Fehse, B. A multicolor panel of novel lentiviral ‘gene ontology’ (LeGO) vectors for functional gene analysis. *Mol. Ther.* **16,** 698–706 (2008).

133. Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* **32,** 1262–7 (2014).

134. Sankaran, V. G. *et al.* Human fetal hemoglobin expression is regulated by the

developmental stage-specific repressor BCL11A. *Science* **322,** 1839–1842 (2008).

135. Sankaran, V. G. *et al.* Developmental and species-divergent globin switching are driven by BCL11A. *Nature* **460,** 1093–1097 (2009).

136. Xu, J. *et al.* Correction of Sickle Cell Disease in Adult Mice by Interference with Fetal Hemoglobin Silencing. *Science* **993,** 993–6 (2011).

137. Hardison, R. C. & Blobel, G. A. GWAS to therapy by genome edits? *Science* **342,** 206–7 (2013).

138. Mandal, P. K. *et al.* Efficient Ablation of Genes in Human Hematopoietic Stem and Effector Cells using CRISPR/Cas9. *Cell Stem Cell* **15,** 643–652 (2014).

139. Liu, P. *et al.* Bcl11a is essential for normal lymphoid development. *Nat. Immunol.* **4,** 525–32 (2003).

140. John, A. *et al.* Bcl11a is required for neuronal morphogenesis and sensory circuit formation in dorsal spinal cord development. *Development* **139,** 1831–41 (2012).

141. Yu, Y. *et al.* Bcl11a is essential for lymphoid development and negatively regulates p53. *J. Exp. Med.* **209,** 2467–2483 (2012).

142. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered and improved PAM specificities. *Nature* **523,** 481–5 (2015).

143. Findlay, G. M., Boyle, E. a., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513,** 120–3 (2014).

144. Bauer, D. E., Kamran, S. C. & Orkin, S. H. Reawakening fetal hemoglobin: Prospects for new therapies for the beta-globin disorders. *Blood* **120,** 2945–2953 (2012).

145. Basak, A. *et al.* Persistence of fetal hemoglobin and altered neurodevelopment due to BCL11A deletions. *JCI* **In press,** (2015).

146. Funnell, A. P. W. *et al.* 2p15-p16.1 microdeletions encompassing and proximal to BCL11A are associated with elevated HbF in addition to neurologic impairment. *Blood* In Press (2015).

147. Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527,** 192–7 (2015).

148. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523,** 481–5 (2015).

149. Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR–Cas9. *Nat. Rev. Genet.* (2015). doi:10.1038/nrg3899

150. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

151. Pinello, L., Xu, J., Orkin, S. H. & Yuan, G.-C. Analysis of chromatin-state plasticity identifies cell-type-specific regulators of H3K27me3 patterns. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E344–53 (2014).

152. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **10,** 957–63 (2013).

153. Kleinstiver, B. P. *et al.* Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. *Nat. Biotechnol.* **33,** 1293–1298 (2015).

154. Uda, M. *et al.* Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 1620–5 (2008).

155. Lettre, G. *et al.* DNA polymorphisms at the BCL11A, HBS1L-MYB, and Beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 11869–11874 (2008).

156. Thein, S. L. *et al.* Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 11346–11351 (2007).

157. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* **42,** 1049–51 (2010).

158. Farrell, J. J. *et al.* A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression. *Blood* **117,** 4935–45 (2011).

159. Stadhouders, R. *et al.* HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *J. Clin. Invest.* **124,** 1699–1710 (2014).

160. Mtatiro, S. N. *et al.* Genome wide association study of fetal hemoglobin in sickle cell anemia in Tanzania. *PLoS One* **9,** e111464 (2014).

161. Bae, H. T. *et al.* Meta-analysis of 2040 sickle cell anemia patients: BCL11A and HBS1L-MYB are the major modifiers of HbF in African Americans. *Blood* **120,** 1961–2 (2012).

162. Ganesh, S. K. *et al.* Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat. Genet.* **41,** 1191–8 (2009).

163. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.* **41,** 1182–1190 (2009).

164. Kamatani, Y. *et al.* Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat. Genet.* **42,** 210–215 (2010).

165. Menzel, S., Garner, C., Rooks, H., Spector, T. D. & Thein, S. L. HbA2 levels in normal adults are influenced by two distinct genetic mechanisms. *Br. J. Haematol.* **160,** 101–5 (2013).

166. van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492,** 369–75 (2012).

167. Chen, Z. *et al.* Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. *Hum. Mol. Genet.* **22,** 2529–38 (2013).

168. Andreani, M. *et al.* Quantitatively different red cell/nucleated cell chimerism in patients with long-term, persistent hematopoietic mixed chimerism after bone marrow transplantation for thalassemia major or sickle cell disease. *Haematologica* **96,** 128–133 (2011).

169. Chang, J. C., Ye, L. & Kan, Y. W. Correction of the sickle cell mutation in embryonic stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 1036–1040 (2006).

170. Firth, A. L. *et al.* Functional Gene Correction for Cystic Fibrosis in Lung Epithelial Cells Generated from Patient iPSCs. *Cell Rep.* **12,** 1385–1390 (2015).

171. Lokody, I. Genetic therapies: Correcting genetic defects with CRISPR–Cas9. *Nat. Rev. Genet.* **15,** 63–63 (2013).

172. Long, C. *et al.* Prevention of muscular dystrophy in mice by CRISPR/Cas9-mediated editing of germline DNA. *Science* **345,** 1184–8 (2014).

173. Ousterout, D. G. *et al.* Multiplex CRISPR/Cas9-based genome editing for correction of dystrophin mutations that cause Duchenne muscular dystrophy. *Nat. Commun.* **6,** 6244 (2015).

174. Wu, Y. *et al.* Correction of a genetic disease by CRISPR-Cas9-mediated gene editing in mouse spermatogonial stem cells. *Cell Res.* **25,** 67–79 (2015).

175. Osborn, M. J. *et al.* Fanconi Anemia Gene Editing by the CRISPR/Cas9 System. *Hum. Gene Ther.* **26,** 114–126 (2015).

176. Chang, C.-W. *et al.* Modeling Human Severe Combined Immunodeficiency and Correction by CRISPR/Cas9-Enhanced Gene Targeting. *Cell Rep.* **12,** 1668–1677 (2015).

177. Flynn, R. *et al.* CRISPR-mediated genotypic and phenotypic correction of a chronic granulomatous disease mutation in human iPS cells. *Exp. Hematol.* **43,** 838–848.e3 (2015).

178. Hoban, M. D. *et al.* Correction of the sickle cell disease mutation in human hematopoietic stem/progenitor cells. *Blood* **125,** 2597–2605 (2015).

179. Zou, J., Mali, P., Huang, X., Dowey, S. N. & Cheng, L. Site-specific gene correction of a point mutation in human iPS cells derived from an adult patient with sickle cell disease. *Blood* **118,** 4599–4608 (2011).

180. Sun, N. & Zhao, H. Seamless correction of the sickle cell disease mutation of the *HBB* gene in human induced pluripotent stem cells using TALENs. *Biotechnol. Bioeng.* **111,** 1048–1053 (2014).

181. Xie, F. *et al.* Seamless gene correction of -thalassemia mutations in patient-specific iPSCs using CRISPR/Cas9 and piggyBac. *Genome Res.* **24,** 1526–33 (2014).

182. Wang, J. *et al.* Homology-driven genome editing in hematopoietic stem and progenitor cells using ZFN mRNA and AAV6 donors. *Nat. Biotechnol.* (2015). doi:10.1038/nbt.3408

183. Boissel, S. *et al.* megaTALs: a rare-cleaving nuclease architecture for therapeutic genome engineering. *Nucleic Acids Res.* **42,** 2591–2601 (2014).

184. Sather, B. D. *et al.* Efficient modification of CCR5 in primary human hematopoietic cells using a megaTAL nuclease and AAV donor template. *Sci. Transl. Med.* **7,** 307ra156 (2015).

185. Yu, C. *et al.* Small Molecules Enhance CRISPR Genome Editing in Pluripotent Stem Cells. *Cell Stem Cell* **16,** 142–147 (2015).

186. Pinder, J., Salsman, J. & Dellaire, G. Nuclear domain 'knock-in' screen for the evaluation and identification of small molecule enhancers of CRISPR-based genome editing. *Nucleic Acids Res.* **43,** 9379–92 (2015).

187. Lin, S., Staahl, B., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *Elife* **3,** 1–13 (2014).

188. Orthwein, A. *et al.* A mechanism for the suppression of homologous recombination in G1 cells. *Nature* **528,** 422–426 (2015).

189. Boitano, A. E. *et al.* Aryl Hydrocarbon Receptor Antagonists Promote the Expansion of Human Hematopoietic Stem Cells. *Science* **329,** 1345–1348 (2010).

190. Wagner, J. E. *et al.* Phase I/II Trial of StemRegenin-1 Expanded Umbilical Cord Blood Hematopoietic Stem Cells Supports Testing as a Stand-Alone Graft. *Cell Stem Cell* (2015). doi:10.1016/j.stem.2015.10.004

191. Fares, I. *et al.* Pyrimidoindole derivatives are agonists of human hematopoietic stem cell self-renewal. *Science* **345,** 1509–1512 (2014).

192. Menzel, S. *et al.* A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat. Genet.* **39,** 1197–9 (2007).

193. Nuinoon, M. *et al.* A genome-wide association identified the common genetic variants influence disease severity in β0-thalassemia/hemoglobin E. *Hum. Genet.* **127,** 303–314 (2010).

194. Solovieff, N. *et al.* Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5' olfactory receptor gene cluster. *Blood* **115,** 1815–1822 (2010).

195. Bhatnagar, P. *et al.* Genome-wide association study identifies genetic variants influencing F-cell levels in sickle-cell patients. *J. Hum. Genet.* **56,** 316–323 (2011).

196. Sankaran, V. G. *et al.* A functional element necessary for fetal hemoglobin silencing. *N. Engl. J. Med.* **365,** 807–14 (2011).

197. Sankaran, V. G. & Orkin, S. H. The switch from fetal to adult hemoglobin. *Cold Spring Harb. Perspect. Med.* **3,** a011643 (2013).

198. Bank, A. Regulation of human fetal hemoglobin: new players, new complexities. *Blood* **107,** 435–443 (2006).

199. Forget, B. G. Molecular basis of hereditary persistence of fetal hemoglobin. *Ann. N. Y. Acad. Sci.* **850,** 38–44 (1998).

200. Chakalova, L. *et al.* The Corfu δβ thalassemia deletion disrupts γ-globin gene silencing and reveals post-transcriptional regulation of HbF expression. *Blood* **105,** 2154–2160 (2005).

201. Comi, P. *et al.* Globin chain synthesis in single erythroid bursts from cord blood: studies on gamma leads to beta and G gamma leads to A gamma switches. *Proc. Natl. Acad. Sci. U. S. A.* **77,** 362–5 (1980).

202. Ottolenghi, S. *et al.* Sardinian G gamma-HPFH: a T----C substitution in a conserved 'octamer' sequence in the G gamma-globin promoter. *Blood* **71,** 815–817 (1988).

203. Martin, D., Tsai, S.-F. & Orkin, S. H. Increased γ-globin expression in a nondeletion HPFH mediated by an erythroid-specific DNA-binding factor. *Nature* **338,** 435–438 (1989).

204. Wienert, B. *et al.* Editing the genome to introduce a beneficial naturally occurring mutation associated with increased fetal globin. *Nat. Commun.* **6,** 7085 (2015).

205. Traxler, E. *et al.* Genome Editing Recreates Hereditary Persistence of Fetal Hemoglobin in Primary Human Erythroblasts (Abstract 640). in 2015 American Society of Hematology Annual Meeting

206. Funnell, A. P. W. *et al.* 2p15-p16.1 microdeletions encompassing and proximal to BCL11A are associated with elevated HbF in addition to neurologic impairment. *Blood* **126,** 89–94 (2015).

207. Basak, A. *et al.* BCL11A deletions result in fetal hemoglobin persistence and neurodevelopmental alterations. *J. Clin. Invest.* **125,** 2363–8 (2015).

208. Kuo, T. Y., Chen, C. Y. & Hsueh, Y. P. Bcl11A/CTIP1 mediates the effect of the glutamate receptor on axon branching and dendrite outgrowth. *J. Neurochem.* **114,** 1381–1392 (2010).

209. Benitez, C. M. *et al.* An Integrated Cell Purification and Genomics Strategy Reveals Multiple Regulators of Pancreas Development. *PLoS Genet.* **10,** e1004645 (2014).

210. Khaled, W. T. *et al.* BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. *Nat. Communic.* **6,** 5987 (2015).

211. Yu, Y. *et al.* Bcl11a is essential for lymphoid development and negatively regulates p53. *J. Exp. Med.* **209,** 2467–83 (2012).

212. Lin, Y., Zhang, Q., Zhang, H., Liu, W. & Liu, C. Transcription factor and miRNA co-regulatory network reveals shared and specific regulators in the development of B cell

and T cell. *Sci. Rep.* **5,** 15215 (2015).

213. Powers, A. N. & Satija, R. Single-cell analysis reveals key roles for Bcl11a in regulating stem cell fate decisions. *Genome Biol.* 1186–1187 (2015). doi:10.1186/s13059-015-0778-y

214. Tsang, J. C. H. *et al.* Single-cell transcriptomic reconstruction reveals cell cycle and multi-lineage differentiation defects in Bcl11a-deficient hematopoietic stem cells. *Genome Biol.* **16,** 178 (2015).

215. Guda, S. *et al.* miRNA-embedded shRNAs for lineage-specific BCL11A knockdown and hemoglobin F induction. *Mol. Ther.* **23,** 1465–74 (2015).

216. Howe, S. J. *et al.* Insertional mutagenesis in combination with acquired somatic mutations leads to leukemogenesis following gene therapy of SCID-X1. *J. Clin* **118,** 3143–50 (2008).

217. Vierstra, J. *et al.* Functional footprinting of regulatory DNA. *Nat. Methods* **12,** 927–930 (2015).

218. Masuda, T. *et al.* Transcription factors LRF and BCL11A independently repress expression of fetal hemoglobin. *Science* **351,** 285–289 (2016).

219. Maeda, T. *et al.* LRF is an essential downstream target of GATA1 in erythroid development and regulates BIM-dependent apoptosis. *Dev. Cell* **17,** 527–40 (2009).

220. Lunardi, A., Guarnerio, J., Wang, G., Maeda, T. & Pandolfi, P. P. Role of LRF/Pokemon in lineage fate decisions. *Blood* **121,** 2845–53 (2013).

221. Musallam, K. M. *et al.* Fetal hemoglobin levels and morbidity in untransfused patients with beta-thalassemia intermedia. *Blood* **119,** 364–367 (2012).

222. Wilber, A. *et al.* Therapeutic levels of fetal hemoglobin in erythroid progeny of β-thalassemic CD34+ cells after lentiviral vector-mediated gene transfer. *Blood* **117,** 2817–26 (2011).

223. Mettananda, S., Gibbons, R. J. & Higgs, D. R. α-Globin as a molecular target in the

treatment of β-thalassemia. **125,** 3694–3702 (2015).

224. Kan, Y. W. & Nathan, D. G. Mild thalassemia: the result of interactions of alpha and beta thalassemia genes. *J. Clin. Invest.* **49,** 635–642 (1970).

225. Thein, S. L. Genetic modifiers of the β-haemoglobinopathies. *Br. J. Haematol.* **141,** 357–366 (2008).

226. Renneville, A. *et al.* EHMT1 and EHMT2 inhibition induces fetal hemoglobin expression. *Blood* **126,** 1930–9 (2015).

227. Krivega, I. *et al.* Inhibition of G9a methyltransferase stimulates fetal hemoglobin production by facilitating LCR/γ-globin looping. *Blood* **126,** 665–673 (2015).

228. Lee, Y. T. *et al.* LIN28B-mediated expression of fetal hemoglobin and production of fetal-like erythrocytes from adult human erythroblasts ex vivo. *Blood* **122,** 1034–41 (2013).

229. Hendel, A. *et al.* Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. *Nat. Biotechnol.* **33,** 985–9 (2015).

230. Nienhuis, A. W. & Persons, D. A. Development of gene therapy for thalassemia. *Cold Spring Harb. Perspect Med* **2,** a011833 (2012).

231. Nienhuis, A. Development of gene therapy for blood disorders: an update. *Blood* **122,** 1556–1564 (2013).

232. Papapetrou, E. P., Zoumbos, N. C. & Athanassiadou, a. Genetic modification of hematopoietic stem cells with nonviral systems: past progress and future prospects. *Gene Ther.* **12,** S118–S130 (2005).

233. Urnov, F. D. *et al.* Clinical-Scale Genome Editing of the Human BCL11A Erythroid Enhancer for Treatment of the Hemoglobinopathies (Abstract 204). in 2015 American Society of Hematology Annual Meeting

234. Mandal, P. K. *et al.* Efficient Ablation of Genes in Human Hematopoietic Stem and Effector Cells using CRISPR/Cas9. *Cell Stem Cell* **15,** 643–652 (2014).

235. Buechele, C. *et al.* MLL leukemia induction by genome editing of human CD34+

hematopoietic cells. *Blood* **126,** 1683–1695 (2015).

236. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33,** 187–197 (2015).

237. Frock, R. L. *et al.* Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat. Biotechnol.* **33,** 179–86 (2015).

238. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* (2015).

239. Kleinstiver, B. P. *et al.* High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* (2016). doi:10.1038/nature16526

240. Yannaki, E. *et al.* Hematopoietic stem cell mobilization for gene therapy: superior mobilization by the combination of granulocyte-colony stimulating factor plus plerixafor in patients with β-thalassemia major. *Hum. Gene Ther.* **24,** 852–60 (2013).

241. Fitzhugh, C. D., Hsieh, M. M., Bolan, C. D., Saenz, C. & Tisdale, J. F. Granulocyte colony-stimulating factor (G-CSF) administration in individuals with sickle cell disease: time for a moratorium? *Cytotherapy* **11,** 464–471 (2009).

242. Yannaki, E. *et al.* Hematopoietic stem cell mobilization for gene therapy of adult patients with severe β-thalassemia: results of clinical trials using G-CSF or plerixafor in splenectomized and nonsplenectomized subjects. *Mol. Ther.* **20,** 230–8 (2012).

243. Karponi, G. *et al.* Plerixafor+G-CSF-mobilized CD34+ cells represent an optimal graft source for thalassemia gene therapy. *Blood* **126,** 616–620 (2015).

244. Ellis, E. L. & Delbrück, M. The Growth of Bacteriophage. *J. Gen. Physiol.* **22,** 365–384 (1939).

245. Stent, G. *Molecular Biology of Bacterial Viruses*. (W. H. Freeman and Company, 1963).

246. Choi, C., Kuatsjah, E., Wu, E. & Yuan, S. The Effect of Cell Size on the Burst Size of T4 Bacteriophage Infections of Escherichia coli B23. *J. Exp. Microbiol. Immunol.* **14,** 85–91 (2010).

247. Cui, F., Sirotin, M. V & Zhurkin, V. B. Impact of Alu repeats on the evolution of human p53 binding sites. *Biol. Direct* **6,** 2 (2011).

248. Crocker, J. *et al.* Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness. *Cell* 191–203 (2015). doi:10.1016/j.cell.2014.11.041